STUDENT TEACHER INTERACTION ANALYSIS WITH EMOTION RECOGNITION FROM VIDEO AND AUDIO INPUT: RESEARCH

Rajat Dubey[1], Vedant Juikar[2], Roshani Surwade[3], Prof. Rasika Shintre[4]

[1,2,3]B.E. Student, Computer Department,
[4]Project Guide, Smt. Indira Gandhi College of Engineering Navi Mumbai, Maharashtra, India

Abstract - *Emotions are significant because they are essential to the learning process. Emotions, behaviour, and thoughts are intimately connected in such a way that the sum of these three factors determines how we act and what choices we make. The selection of a database, identifying numerous speech-related variables, and making an appropriate classification model choice are the three key hurdles in emotion recognition. In order to understand how these emotional states relate to students' and teachers' comprehension, it is important to first define the facial physical behaviours that are associated with various emotional states. The usefulness of facial expression and voice recognition between a teacher and student in a classroom was first examined in this study.*

Key words: Speech Emotion Recognition, Facial Emotion Recognition, JAVA, Node.js, Computer Neural Network, Recurrent Neural Network, MFCC, Support Vector Machine

1. INTRODUCTION

In artificial intelligence, it is common practice to take real time photos, videos, or audios of people in order to analyze their facial and verbal expressions minutely. Because there is very little facial muscle twisting, it is difficult for machines to recognize emotions, which leads to inconsistent results. The contact between teachers and students is the most important component of any classroom setting. In interactions between teachers and pupils, the impact brought about by facial expressions and voice is very strong.

Fig. 1 Basic Emotion Recognition

The key sources of information for figuring out a person's interior feelings in humans are speech and facial expressions. Real-world classrooms allow for in-person interactions. In order to accomplish this, many schools have regularly scheduled chat rooms with audio and video conferencing interactions. In these spaces, students may easily connect with one another and the instructor just as they would in a traditional classroom. Through this technique, the teacher will be able to identify the student's facial expressions or spoken expressions of satisfaction with me as they occur throughout interactions with the student. The fundamental tenet is that teachers must be able to read students' minds and observe their facial expressions.

We analyzed at whether the students' facial action units conveyed their emotions in relation to comprehension. The primary hypothesis of the first step of this study proposed that students frequently use nonverbal communication in the classroom, and that this nonverbal communication takes the form of emotion recognition from video and audio. This, in turn, enables lecturers to gauge the level of understanding of the students.

The models that are available include Voice Emotion Recognition, which recognizes emotions through speech but is unable to produce results from video input. Face Emotion Recognition uses video to identify emotions, however it is unable to do so with audio input, and no analysis report is produced.

2. LITERATURE SURVEY

We provide an overview of some current research in the fields of speech emotion recognition (SER) and facial expression recognition (FER), as well as some FER systems utilized in classrooms as the foundation for teacher-student interaction, because the face plays a significant role in the expression and perception of emotions.

A system that allowed them to collect audio recordings of the emotions of irritation, happiness, and melancholy together with their STE, pitch, and MFCC coefficients. Only the three basic emotions—anger, happiness, and sadness—were identified. Using feature vectors as input, the multi-class Support vector machine (SVM) generates a model for each emotion. Deep Belief Networks (DBNs) have an accuracy rate that is roughly 5% greater than Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) when compared to them. The results demonstrate how much better the features recovered using Deep Belief Networks are than the original feature.

The speech signal's audio feature is the sound. Feature extraction is the process of extracting a little amount of information from vocal expression that can then be utilized to act for each speaker. Feature extraction and feature classification must be the primary SER strategies. For feature classification, both linear and nonlinear classifiers can be utilized. Support Vector Machines are a popular classifier in linear classifiers (SVMs). These kinds of classifiers are useful for SER since speech signals are thought to vary. Deep learning approaches have more benefits for SER than conventional approaches. Deep learning approaches have the capacity to recognize complicated internal structure and do not require manual feature extraction or tweaking. The majority of the current face recognition algorithms fall into one of two categories: geometric feature-based or image template-based. While local face characteristics are extracted and their geometric and aesthetic properties are used in feature-based approaches. Our primary point of emphasis during social interactions is the face, which is crucial for expressing identity and emotion. We can quickly recognise familiar faces even after being apart for a long period because we have learned to recognise hundreds of faces over the course of our existence. Particularly intriguing are computational models for face recognition since they can advance both theoretical understanding and real-world use. System that allows computers to categorize numerous audio voice files into distinct emotions including joyful, sad, angry, and neutral. In this paper, we used Tennent Meeting as an example for mode testing. The framework primarily comprises of two parts: online course platforms and a deep learning model based on CNN.

3. METHODOLOGY

Recognizing speech and facial emotions relies heavily onfacial communication and expression. They are based on the physical and psychological circumstances.

A dataset is trained and tested as part of the supervised learning process used to train the emotion recognition system. Face Detection, Image Pre-processing, Feature Extraction, and Classification are all steps in the basic process of a facial emotion identification system.

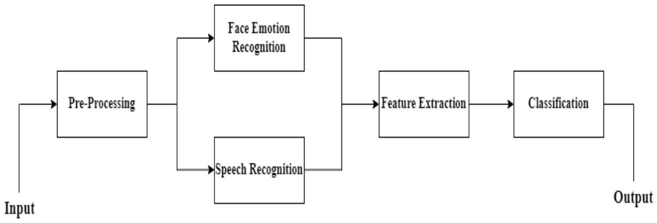


Figure 2. Emotion Recognition System Block Diagram

3.1 Face Emotion Recognition:

Tracking the face through the unprocessed input photos is what it entails. On the training dataset, OpenCV is used to implement it after CNN processes it. There are continually new models based on the CNN structure that have produced better outcomes for the identification of facial emotions.

3.1 Speech Recognition:

Two different types of models—the discrete speech emotion model and the continuous speech emotion model—are included in the SER model. The first model displays a variety of individualistic emotions, suggesting that a particular voice has only one individualistic emotion, whereas the second model indicates that the emotion is in the emotion space and that each emotion has its own special strength. It makes use of a range of emotions, including neutral, disgust, wrath, fear, surprise, joy, happiness, and sadness.

3.2 Feature Extraction:

To identify emotions, we have employed face traits. Finding and extracting appropriate features is one of the most important components of an emotion recognition system. These elements were selected to represent the information that was wanted. After pre-processing, facial features with high expression intensity are retrieved from an image, such as the corners of the mouth, nose, forehead creases, and eyebrows.

Different aspects based on speech and facial expression are combined. In comparison to systems created utilising individual features, several studies on the combination of features have shown a significant improvement in classification accuracy.

3.3 Classification:

A classification technique is employed to minimise the data dimensionality because the data obtained from the extraction of voice and facial features has a very high dimensionality. Support vector machine algorithm is used in this procedure. To recognise various patterns, SVM is employed. SVM is employed to train the proper feature based data set, and even with the availability of only a modest amount of training data, it produces high classification accuracy.

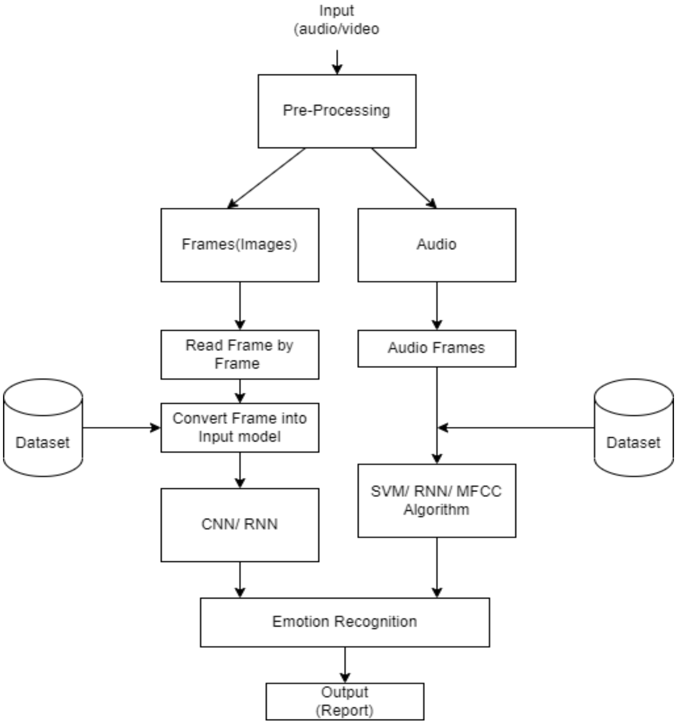


Fig. 3 Architecture Of Emotion Recognition

4. ALGORITHM FOR EMOTION RECOGNITION

4.1 Support Vector Machines (SVM)

Data analysis and classification algorithms under supervision are effective at classifying students' facial expressions.

An algorithm for machine learning that can be used for classification is support vector machines.

The most often used and frequently most effective algorithms for voice emotion recognition are Support Vector Machines (SVM) with non-linear kernels. Using a kernel mapping function, an SVM with a non-linear kernel converts the input feature vectors into a higher dimensional feature space. Classifiers that are non-linear in the original space can be made linear in the feature space by using the proper nonlinear kernel functions.

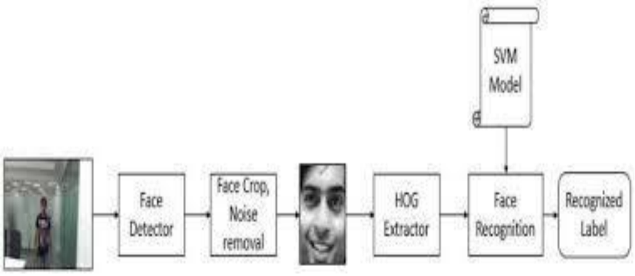


Fig. 4 SVM in Face Recognition

Multiple photographs of the same face taken at various facial expressions will yield a variety of sample images. The image can be cropped and saved as a sample image for examination when the face has been located. Face Detection, Face Prediction, and Face Tracking are the three main steps in the process of identifying a face in a video sequence.

Face recognition also involves the tasks carried out by the Face Capture program. An HOG face highlight vector must be eliminated in order to perceive the obtained face. After that, this vector is used in the SVM model to determine a coordination score for the information vector containing all of the names. The SVM restores the mark with the highest score, which testifies to the certainty of the nearest coordinate within the prepared face information. A program was also created to do SVM-based recognition of spontaneous facial expressions in the video, which was speculatively mentioned as an addition in the proposal.

4.3 Mel-Frequency Cepstrum Coefficient (MFCC)

The preferred method of spectral property encoding for voice communications. These are the greatest because they take into account how sensitively humans perceive frequencies, making them the best for speech recognition.

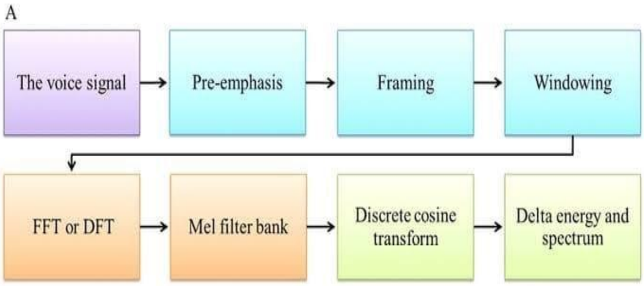


Fig. 5 MFCC in Speech Recognition

The task of supervised learning is speech recognition. The audio signal will be the input for the speech recognition issue, and we must predict the text from the audio signal. Since there would be a lot of noise in the audio signal, we cannot feed the raw audio signal into our model. It has been shown that using the base model's input as a feature extracted input instead of the input's raw audio signal would result in significantly higher performance.

4.4 Convolutional Neural Networks (CNN)

The applied deep learning model's architecture, which preferred the FER above other comparable models, has also been established. Following a convolutional layer with 32 feature maps, the input layer is followed by two blocks that each contain two convolutional layers and one max-pooling layer with 64 feature maps. The first convolutional layer's kernel size is set to 3 3, the second's to 5 5, the max-pooling layers each have a kernel of size 2 2 and stride 2, and as a result, the input image will be compressed to a quarter of its original size. Rectified Linear Units are used as the activation function in the 2 subsequent fully connected layers, A Dropout is added after each of the two completely connected layers to prevent over-fitting; in this paper, the two values are both set to 0.5. The next output layer consists of 8 units, and softmax is used as the activation function to categorize the expressions looked at as contempt, anger, disgust, fear, happiness, sorrow, and neutral.

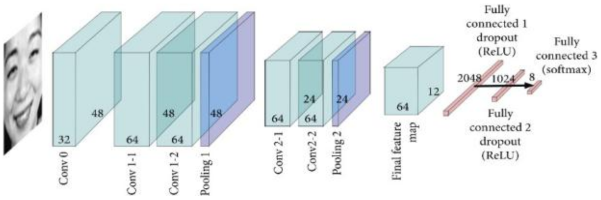


Fig. 6 CNN model in emotion recognition

I. SSD MobileNet V1

Use of SSD Mobilenet V1 for face detection. It is to create a face recognition system, which entails developing a face detector to determine the location of a face in an image and a face identification model to identify whose face it is by comparing it to the already existing database of faces. In order to perform tasks like object detection and picture recognition on mobile devices with limited computational resources, the SSD MobileNet V1 convolutional neural network architecture was developed.

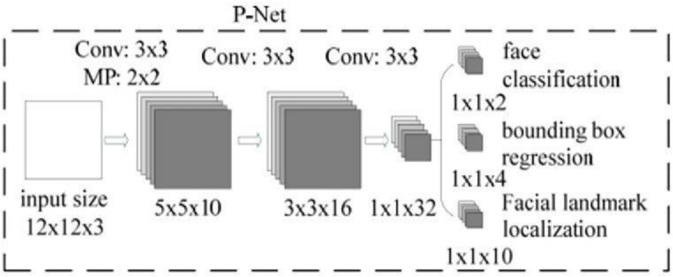


Fig. 7 CNN model in emotion recognition

This project implements a SSD (Single Shot Multibox Detector) based on MobileNetV1. The neural net will compute the locations of each face in an image and will return the bounding boxes together with its probability for each face. This face detector is aiming towards obtaining high accuracy in detecting face bounding boxes instead of low inference time.

II. Tiny Face Detector

The Tiny Face Detector is a real time face detector, which is much faster, smaller and less resource consuming compared to the SSD Mobilenet V1 face detector, in return it performs slightly less well on detecting small faces. The face detector has been trained on a custom dataset of ~14K images labeled with bounding boxes. Furthermore the model has been trained to predict bounding boxes, which entirely cover facial feature points, thus it general produces better results.

III. MTCNN

MTCNN should be able to detect a wide range of face bounding box sizes. MTCNN is a 3 stage cascaded CNN, which simultaneously returns 5 face landmark points along with the bounding boxes.

We propose a deep cascaded multi-task framework which exploits the inherent correlation between detection and alignment to boost up their performance. In particular, our framework leverages a cascaded architecture with three stages of carefully designed deep convolutional networks to predict face and landmark location in a coarse-to-fine manner. In addition, we propose a new online hard sample mining strategy that further improves the performance in practice. Our method achieves superior accuracy over the state-of-the-art techniques on the challenging FDDB and WIDER FACE benchmarks for face detection, and AFLW benchmark for face alignment, while keeps real time performance.

5. PERFORMANCE MATRIX

Convolutional neural network (CNN) analysis of audio and video often results in the creation of a performance matrix that includes an assessment of the CNN model's precision and efficiency in processing audio and video input. Following measures shows the performance of model.

a. **Precision:** Precision measures the ratio of real positive predictions—i.e., outcomes that were accurately forecasted as positive—to the total number of positive predictions which CNN model made.

b. **Recall:** Recall, also referred to as sensitivity or the true positive rate, calculates the ratio of true positive forecasts to all of the data's actual positive outcomes.

c. **F1-score:** The harmonic mean of accuracy and memory, which offers a balanced assessment of both precision and recall, is the F1-score. It is a frequently employed metric where recall and precision are of equal importance. In Fig. F1 score of different algorithms is shown.

d. **Accuracy:** This gauges how accurately the CNN model has predicted the outcomes for the analysed audio and video data overall. It is determined as the proportion of accurately predicted results to the entire sample size.

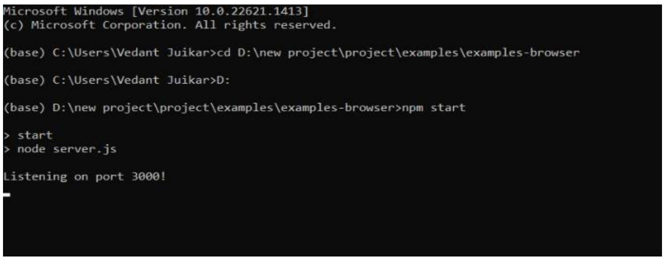
	Video (CNN)			Audio	
Measure/Classifier	Tiny face Detector	SSD MobileNet V1	MTCNN	MFCC	SVM
Precision	0.87	0.98	0.99	0.89	0.92
Recall	0.98	0.97	0.98	0.78	0.91
F1 Score	0.96	0.95	0.94	0.88	0.86
Accuracy	0.97	0.96	0.95	0.89	0.90

Table. 1 Performance Matrix

6. IMPLEMENTATION

6.1 Input

Launch the Anaconda software, navigate to the Node.js file in the command prompt, and launch the programme file for video emotion identification. The programme began to function, and using the audio and video input, it recognised the emotions and produced the graph of that feeling for the teacher and the pupil.



```
Microsoft Windows [Version 10.0.22621.1413]
(c) Microsoft Corporation. All rights reserved.

(base) C:\Users\Vedant Juikar>cd D:\new project\project\examples\examples-browser

(base) C:\Users\Vedant Juikar>D:

(base) D:\new project\project\examples\examples-browser>npm start

> start
> node server.js

Listening on port 3000!
```

Fig. 8 Initial command prompt

6.2 Audio/Video Detection

Choose the interaction analysis video between the teacher and students. Play the video. Two different emotion kinds are recognised simultaneously: face emotion and spoken emotion detection using video input. The UI created for emotion detection counts the amount of facial expressions made by the pupils in the video and while engaging with the teacher, therefore their audio will be counted as well.

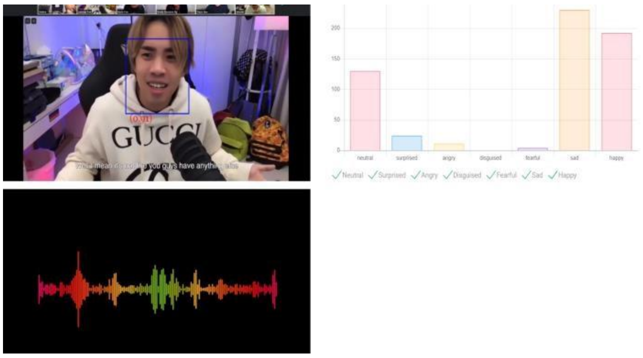


Fig. 9. Audio and facial Recognition

6.3 Emotion Graph

The graph is created after playing a video. The created graph is based on the audio and video's recognized emotions. The total amount of emotions based on the audio and video are displayed as a graph.

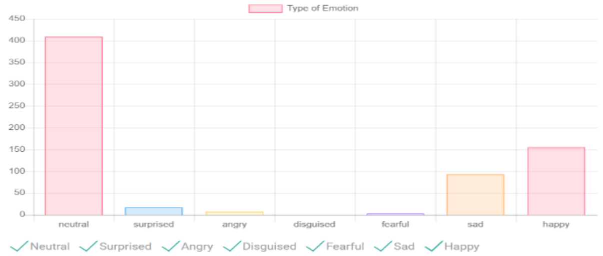
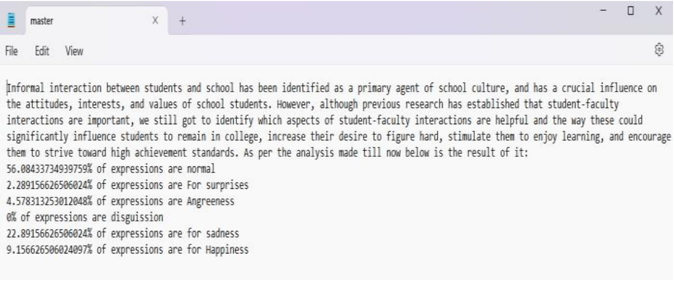


Fig. 10 Emotion Graph

6.4 Audio/ Video Report

As a video is played, emotions are picked up, and we created a graph to analyse those feelings in relation to the teacher-student interaction. There are many levels of emotion seen in the created graph.



```
Informal interaction between students and school has been identified as a primary agent of school culture, and has a crucial influence on
the attitudes, interests, and values of school students. However, although previous research has established that student-faculty
interactions are important, we still got to identify which aspects of student-faculty interactions are helpful and the way these could
significantly influence students to remain in college, increase their desire to figure hard, stimulate them to enjoy learning, and encourage
them to strive toward high achievement standards. As per the analysis made till now below is the result of it:
56.08433734939759% of expressions are normal
2.289156626506024% of expressions are For surprises
4.578313253012048% of expressions are Angreeness
0% of expressions are disguission
22.89156626506024% of expressions are for sadness
9.156626506024097% of expressions are for Happiness
```

Fig. 11 Audio-Video Report

7. RESULT ANALYSIS

This proposal is mainly used for the recorded teacher-student interaction in a classroom to identify the emotions of the teacher as well as the student through the audio and video input. It is a way of generating a report on the range of emotions of the teacher and student from the facial and audio frames. With this feature, the emotions of the students are identified. This helps out the lecture to make a lecture more interactive, increase their desire to figure hard, stimulate them to enjoy learning and encouragement them to strive towards high achievement standard.

8. CONCLUSION

The analysis of audio and facial emotions holds great promise in enhancing our comprehension of human emotions and conduct. By incorporating data from both modalities, the precision and dependability of emotion recognition algorithms can be enhanced, and fresh applications of this technology are predicted to surface in domains like healthcare, education, and entertainment. Moreover, there is a need to ensure that the data used for training and testing emotion recognition algorithms is diverse and representative of the populationthe future of audio and facial emotion analysis is exciting and holds immense potential to improve our understanding of human emotions and behavior. With further advancements in technology and research, we can expect to see new and innovative applications of this technology in a wide range of domains, making our interactions with technology and with each other more efficient and empathetic.

9. FUTURE SCOPE

The future scope of audio and facial emotion analysis is vast and promising. With the increasing development of machine learning and artificial intelligence, it is expected that the accuracy and reliability of emotion detection algorithms will continue to improve. In addition, new applications of this technology are likely to emerge in fields such as healthcare, education, and entertainment. For example, in healthcare, emotion detection could be used to monitor patients for signs of mental health issues, such as depression or anxiety. In education, it could be used to analyze student engagement and identify areas for improvement in teaching methods.

10. REFERENCES

[1] G. Tonguç and B. O. Ozkara, "Automatic recognition of student emotions from facial expressions during a lecture," Computers & Education, vol. 148, Article ID 103797, 2020.

[2] S. Lugović*, I. Dunđer** and M. Horvat "Techniquesand Applications of Emotion Recognition in Speech", May 2016

[3] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 97–115, 2001

[4] Z.Zeng,M.Pantic,G.I.Roisman,andT.S.Huang,"Asurvey of affect recognition methods: audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39– 58, 2009.

[5] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in Advances in Face Detection and Facial Image Analysis, pp. 63–100, Springer, Cham, Switzerland, 2016

BIOGRAPHERS

Rajat Dubey is pursuing the Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi college Of Engineering ,Navi Mumbai

Vedant Juikar is pursuing the Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi college Of Engineering ,Navi Mumbai

Roshani Surwade is pursuing the Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi college Of Engineering ,Navi Mumbai.

PROF. Rasika Shintre, Obtained the Bachelor degree (B.E. Computer) in the year 2011 from Ramrao Adik Institute of Technology (RAIT), Nerul and Master Degree (M.E. Computer)From Bharti Vidyapeeth College Of Engineering, Navi Mumbai.She is Asst. Prof in Smt. Indira Gandhi college Of Engg. Of Mumbai University and having about 11 years of experience. Her area of interest include Data Mining and Information Retrieval.