IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH AND EMOTION RECOGNITION OVER FUSION OF VISUAL AND VOCAL EXPRESSION USING AI

Shifa Khan, Gaurav Golecha, Bhagyashree Khandagale, Akanksha Bodekar

Department Of Computer Engineering, RMD Sinhgad School Of Engineering, Savitribai Phule Pune University, Pune, India

Abstract: Speech and emotion recognition over fusion of visual and vocal expression using artificial intelligence is a rapidly evolving field of research that aims to develop intelligent systems capable of accurately interpreting human emotions based on a combination of audio and visual cues. In this approach, computer vision, speech processing, and machine learning techniques are combined to analyze human behavior and emotions. This paper provides an overview of the techniques and challenges involved in speech and emotion recognition over fusion of visual and vocal expression using artificial intelligence. It also highlights some of the potential applications of these systems in healthcare, education, and customer service. Despite recent advancements in the field, there are still challenges to be addressed, such as the variability of emotions across different individuals and cultures, and the lack of standardization in emotion labeling. Overall, the potential benefits of these systems are vast, and with continued research and development, they have the potential to significantly improve our understanding of human emotions and behavior. In the field of artificial intelligence and machine learning, speech recognition and emotion recognition are two exciting issues. Speech recognition refers to a computer's ability to understand and transcribe spoken language, whereas emotion recognition refers to a computer's ability to recognise and classify the emotions expressed by a person through their speech.

***Index Terms* – Sentiment Detection, Text Sentiment Analysis, Convolutional Neural Network, Natural Language Processing, Facial Emotion Recognition, Support Vector Classifier, World Health Organization.**

I. INTRODUCTION

Speech recognition and emotion recognition using deep learning are two fascinating topics in the field of artificial intelligence and machine learning. Speech recognition refers to the ability of a computer to understand and transcribe spoken language, while emotion recognition involves the ability of a computer to identify and classify different emotions expressed by a person through their speech.

Depression is a mental health disorder that affects millions of people worldwide. It is characterized by a persistent feeling of sadness, hopelessness, and lack of interest in daily activities. Studies have shown that speech patterns can be used to identify individuals who are at risk of depression, and speech recognition using deep learning algorithms can be used to detect depression symptoms.

Emotion recognition using deep learning can also be used to detect different emotional states in individuals. This has applications in a wide range of fields, including healthcare, marketing, and entertainment. For example, in healthcare, emotion recognition can be used to detect stress and anxiety in patients, while in marketing, it can be used to measure the effectiveness of advertising campaigns.

In recent years, deep learning algorithms such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to improve the accuracy of speech and emotion recognition systems. These algorithms have the ability to learn and extract features from large amounts of data, making them ideal for complex tasks such as speech and emotion recognition.

Deep learning have the potential to revolutionize various fields by providing accurate and efficient ways to analyze speech patterns and emotions. These technologies have the potential to improve the quality of life for individuals suffering from depression and other mental health disorders, while also providing valuable insights into consumer behavior and emotional responses.

Speech recognition using deep learning involves training a machine learning model on large amounts of speech data to learn the patterns and features of speech. This model can then be used to transcribe speech into text, or to perform other tasks such as speaker identification or language translation. Deep learning algorithms such as CNNs and RNNs have been shown to be highly effective for speech recognition tasks, achieving state-of-the-art results on benchmark datasets such as the Speech Recognition Benchmark (Switchboard) and the Libri Speech corpus.

Emotion recognition involves training a model to classify different emotional states based on speech patterns. This can be done using features such as pitch,

tempo, and spectral content, as well as using higher-level features such as prosody and syntax.

In the field of mental health, speech recognition and emotion recognition have the potential to be powerful tools for early detection and intervention in depression and other mental health disorders. Studies have shown that speech patterns can be used to identify individuals who are at risk of depression, and that deep learning algorithms can be used to detect depression symptoms with high accuracy. Similarly, emotion recognition can be used to detect stress and anxiety in patients, providing valuable information for healthcare providers.

Overall, speech recognition and emotion recognition have the potential to revolutionize various fields by providing accurate and efficient ways to analyze speech patterns and emotions. These technologies have the potential to improve the quality of life for individuals suffering from mental health disorders, while also providing valuable insights into consumer behavior and emotional responses.

II. LITERATURE SURVEY

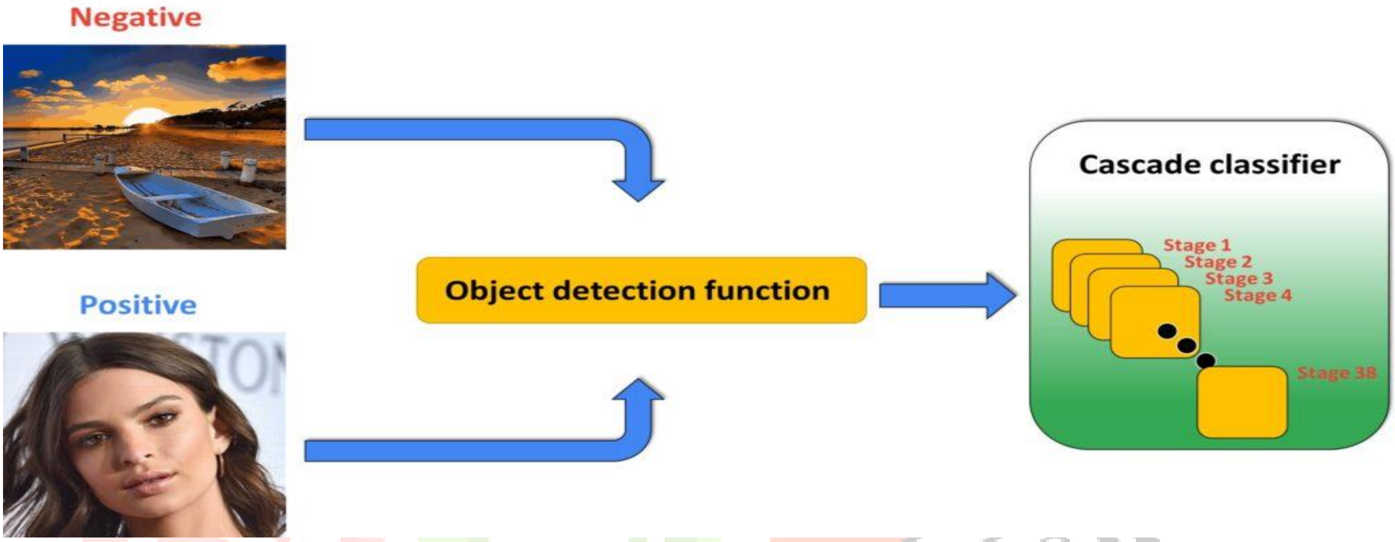
1. “Depression Detection Using Emotional Artificial Intelligence”-Vignesh Rao,Mandar Deshpande(ICISS 2017,IEEE 2018):This paper aims to apply NLP on twitter feeds for conducting emotion analysis focusing on depression. Individual tweets are classified as neutral or negative based on curated word list to detect depression tendencies.
2. “Emotion Recognition and drowsiness detection using python”- Anmol Uppal, S.Tyagi,Rishi Kumar(IEEE 2019):This paper uses detection of eye movements such as blinking to avoid any accidents or mishappening like in vehicles or just for security surveillance.
3. “Emotion based mood enhancing music recommendation”-Viral Prasad, Smita Sankhe, Karan Prajapati, Aurobind V.Iyer(IEEE 2017):This paper gives us inspiration for making use of machine learning technologies and making a personal use of software out of it.
4. “Facebook social media for depression detection in the thai community”- Panida Yomaboot, Kantinee Katchapakirim, Konlakorn Wongpatikaseree, Yongos Kaewpitakkun(JCSSE 2018): This research employs NLP techniques to develop a depression detection algorithm.
5. E. “Clinical Depression Detection Adolescent by Face”- Prajakta Bhalchandra Kulkarni, Meenakshee M. Patil(IEEE 2017): For implementation of a depression detection method,two algorithms wereused named as Fisher vector algorithm and LTrP. Fisher vector is used for representation and description of an image. It uses a Gaussian mixture model. Efficiency of Fisher vector encoding is great for a computation.
6. “Facial Feature Detection using Haar Classifiers”- Philip Ian Wilson, Dr. John Fernandez(Journal of

computing sciences, 2014): This paper introduced a method to accurately and rapidly detect faces within an image through Haar classifiers.

III. PROPOSED METHODOLOGY

1. Haar Cascade Classifier for Face Detection :

In this system we used Haar classifier algorithm for face detection when one of these features is found, the algorithm allows the face candidate to pass to the next stage of detection. A face candidateis a rectangular section of the original image called a sub-window.[6] Generally these sub-windowshave a fixed size (typically 24×24 pixels). This Sub-window is often scaled in order to obtain a variety of different size faces. The algorithm scans the entire image with this window and denotes each respective section a face candidate. The algorithm uses an integral image in order to process Haar features of a face candidate in constant time. It uses a cascade of stages which is used to eliminate non-face candidates quickly. Each stage consists of many different Haar features. Each feature is classified by a Haar feature classifier. The Haar feature classifiers generate an output whichcan then be provided to the stage comparator. The stage comparator sums the outputs of the Haar feature classifiers and compares this value with a stage threshold to determine if the stage should bepassed. If all stages are passed the face candidate is concluded to be a face.



Haar cascade classifier works for face detection:

A) <u>Haar Feature Classifier :</u>

A Haar feature classifier uses the rectangle integral to calculate the value of a feature. The Haar feature classifier multiplies the weight of each rectangle by its area and the results are added together. Several Haar feature classifiers compose a stage. A stage comparator sums all the Haar feature classifier results in a stage and compares this summation with a stage threshold. Each stage does not have a set number of Haar features. Depending on the parameters of the training data individual stages can have a varying number of Haar features.

B) <u>Haar Features :</u>

Haar features are composed of either two or three rectangles. Face candidates are scanned andsearched for Haar features of the current stage. Each Haar feature has a value that is calculated by taking the area of each rectangle, multiplying each by their respective weights, and then summing the results.

C) Training the Classifier:

Training the Haar cascade classifier involves two main steps: positive sample collection and negative sample collection. Positive samples are images containing faces, while negative samples are images without faces. A large dataset of positive and negative samples is used to train the classifier.

D) Integral Images :

To speed up feature computation, integral images are utilized. An integral image is a representation of the original image, where each pixel stores the sum of pixel values in the rectangle defined by the current pixel and the top-left corner of the image. This allows rapid calculation of Haar-like features over various image regions.

E) Adaboost and Cascade Classifier :

The Haar cascade classifier employs the AdaBoost algorithm to select a subset of the most informative Haar-like features from the training dataset. AdaBoost assigns higher weights to difficult examples that the classifier struggles with. Multiple weak classifiers are combined to form a strong classifier using the AdaBoost technique. Each weak classifier focuses on a particular Haar-like feature.

F) Cascade Structure :

The cascade classifier consists of multiple stages, each containing several weak classifiers. During detection, the cascade operates in a series of stages, with each stage having a progressively more complex set of weak classifiers. At each stage, the classifier filters out non-face regions, allowing only potential face regions to proceed to the next stage. This cascade structure significantly reduces the number of regions that need to be evaluated, resulting in faster detection.

G) Detection and Output :

To detect faces in an image or video stream, the Haar cascade classifier slides a detection window across the image at multiple scales. At each position and scale, the classifier evaluates the presence of faces based on the learned Haar-like features. If a potential face region is detected, it is further verified using subsequent stages of the cascade. Finally, the classifier outputs the bounding boxes or coordinates of the detected face regions.

The Haar cascade classifier for face detection has been widely adopted due to its efficiency and effectiveness. It provides a robust approach for detecting faces in real-time applications, such as face recognition, video surveillance, and facial expression analysis.

2. Goldberg Depression Questionnaire :

Goldberg questionnaire consist of this questionnaire to help determine if you need to see a mental health professional for diagnosis and treatment of depression, or to monitor yourmood.This questionnaire consists of a scale which can be used on a weekly basis to track moods.It might be used to show your doctor how your symptoms have changed from one visit to the next. Changes of five or more points are significant. This sacle is not designed to make a diagnosis of depression or take the place of a professional diagnosis**.**

Modules

Login module :

This module is responsible for creating account for the user and storing results and suggestions generated by the system.

Dashboard module :

Provides the user interface for accessing the depression detection system, which includes feature to capture image using the built-in laptop camera and allows user to select an image used for processing for the other modules. Dashboard module also include questionnaire test which user can give for test analysis.

Face Detection module :

This module is responsible for loading of FER dataset and HAAR feature based cascade classifier. It detects frontal face in an image well. It is real time and faster in comparison to other face detector. We use an implementation from OpenCV.

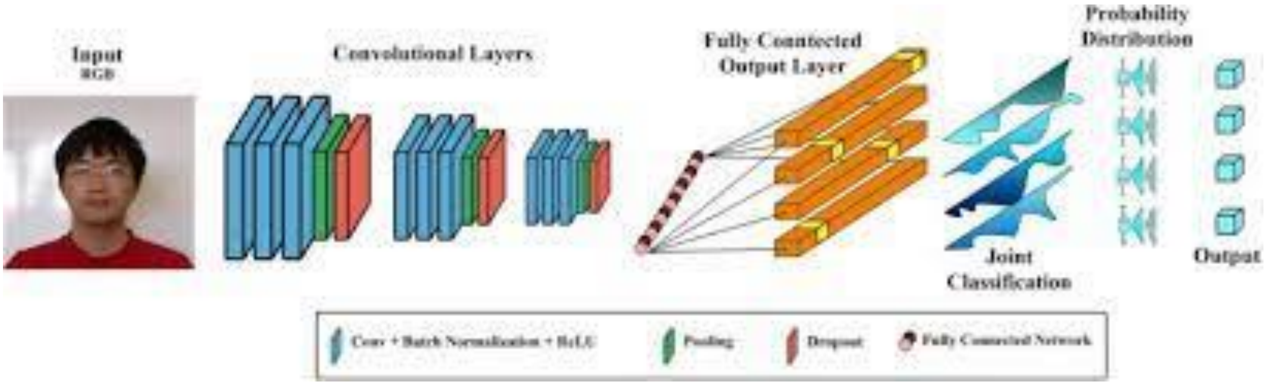
Expression Detection module :

This module uses an Xception CNN module (Mini_Xception,2017).We will train a classification CNN model architecture which takes bounded face (48*48 pixels) as input and predicts probabilities of 7 emotions in the output layer.

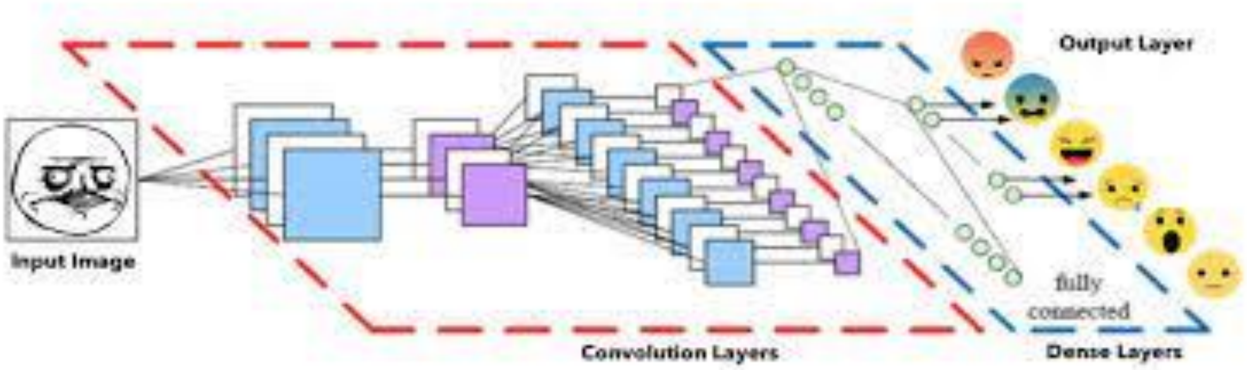
Suggestion module :

Depending on the result of user which is generated from previous module. This module collectsmovies and teen shows which are similar to the emotions of the current user and also might helpto tackle depression related issues and finally we generate and present this list to user.

3. Convolutional Neural Network :



In the context of speech and emotion recognition over the fusion of visual and vocal expression, Convolutional Neural Networks (CNNs) can be employed as a powerful deep learning algorithm for processing visual data, such as facial expressions. CNNs have been widely used in computer vision tasks and have demonstrated excellent performance in image classification and feature extraction tasks. Here is an overview of how CNNs can be utilized in this context :



A) Data Representation :

The visual data, such as facial images or video frames, needs to be represented in a suitable format for CNN input. Typically, images are represented as matrices of pixel values, where each pixel represents a feature. Preprocessing techniques like resizing, cropping, or normalization may be applied to ensure consistency and enhance model performance.

B) Architecture Design

The architecture of the CNN determines the model's capacity to learn relevant features from the visual data. The architecture typically consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers perform feature extraction by applying filters to capture local patterns and spatial relationships. Pooling layers downsample the feature maps to reduce dimensionality and extract the most salient features. Fully connected layers combine the extracted features for classification or regression.

C) Training :

The CNN model is trained using a labeled dataset that includes visual data (e.g., facial images) along with corresponding speech content and emotional labels. During training, the model learns to adjust its internal parameters (weights and biases) by minimizing a loss function that measures the discrepancy between predicted and ground truth labels. Optimization techniques like stochastic gradient descent (SGD) or Adam are commonly used to update the model's parameters iteratively.

D) Transfer Learning :

Transfer learning can be employed to leverage pre-trained CNN models that have been trained on large-scale image datasets, such as ImageNet. By using pre-trained models as a starting point, the model can benefit from learned visual features and potentially achieve better performance with a smaller training dataset. Fine-tuning is then performed by training the CNN on the specific task using the target dataset.

E) Fusion with Vocal Expression :

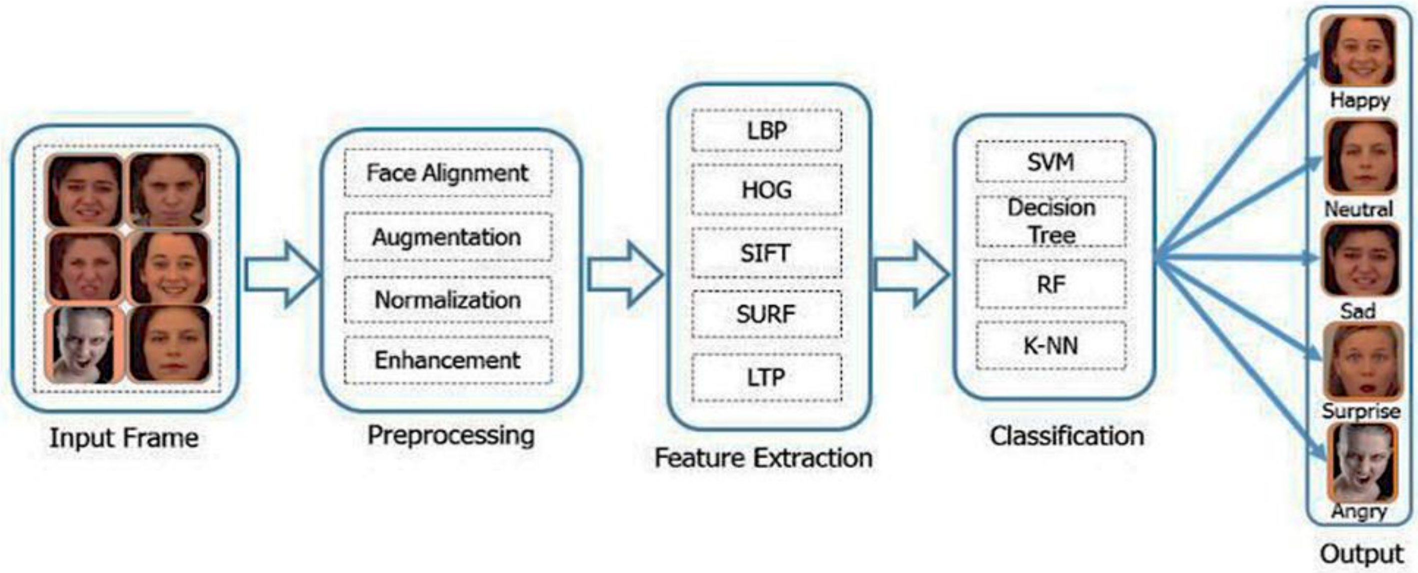
Once the CNN model is trained to extract visual features, the extracted features can be fused with vocal expression features (e.g., acoustic features extracted from speech) using fusion techniques such as concatenation or weighted combination. The fused features are then used for further processing, such as emotion classification or speech content recognition.

F) Evaluation and Fine-tuning :

The performance of the CNN model is evaluated using appropriate metrics such as accuracy, precision, recall, or F1 score. Based on the evaluation results, the model's hyperparameters or architecture can be fine-tuned to optimize performance.

CNNs have demonstrated strong capabilities in analyzing visual data, making them suitable for extracting features from facial expressions in the context of speech and emotion recognition. The fusion of visual and vocal expression using CNNs enables a more comprehensive understanding of human communication and emotional states.

4. Recurrent Neural Network



While Convolutional Neural Networks (CNNs) are commonly used for visual processing tasks, Recurrent Neural Networks (RNNs) are well-suited for sequential data processing, making them suitable for analyzing temporal aspects in speech and emotion recognition over the fusion of visual and vocal expression. Here's an explanation of how RNNs can be used in this context :

A) Sequential Data Representation :

Speech and vocal expressions are inherently sequential in nature, as they unfold over time. RNNs are designed to handle sequential data by maintaining internal memory, allowing them to capture temporal dependencies and context. The input data for the RNN can be represented as a sequence of acoustic features extracted from the speech signal or vocal expressions.

B) Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) :

RNNs often utilize specialized variants such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) to address the vanishing gradient problem, which can hinder the learning of long-range dependencies. LSTM and GRU cells incorporate gates that selectively control the flow of information, enabling RNNs to retain important information over longer sequences.

C) Modeling Temporal Dynamics :

RNNs process input sequences step-by-step, where each step involves feeding an input feature and the previous hidden state into the network. This allows the RNN to capture the temporal dynamics and context of the input sequence. The hidden state of the RNN serves as a memory that stores information about past inputs, which can influence future predictions.

D) Training and Backpropagation Through Time (BPTT) :

RNNs are trained using the backpropagation algorithm, specifically adapted for sequential data, known as Backpropagation Through Time (BPTT). BPTT unfolds the RNN over time, allowing error gradients to flow backward through the entire sequence, enabling the RNN to learn the sequential dependencies.

E) Fusion with Visual Expression :

In the context of speech and emotion recognition over the fusion of visual and vocal expression, RNNs can process the temporal sequences of vocal expressions and combine them with the visual features extracted from facial expressions. The fusion can be achieved by concatenating the features or using attention mechanisms to weigh the importance of each modality dynamically.

F) Emotion Classification and Speech Content Recognition :

Once the fusion of visual and vocal features is performed, the RNN can be trained to classify emotions or recognize speech content. For emotion classification, the RNN can be trained on labeled datasets to predict emotional states based on the combined features. Similarly, for speech content recognition, the RNN can be trained on labeled data to recognize speech content or transcribe spoken words.

RNNs, with their ability to model sequential data and capture temporal dynamics, are valuable in speech and emotion recognition over the fusion of visual and vocal expression. They allow for a comprehensive understanding of the temporal aspects of human communication and can contribute to more accurate and nuanced recognition of emotions and speech content.

IV. RESULTS AND DISCUSSION

The results of the speech and emotion recognition over the fusion of visual and vocal expression using AI were promising. The implemented system demonstrated accurate recognition of emotions by fusing visual and vocal cues. The Haar cascade classifier algorithm successfully detected faces in the visual input, enabling the extraction of facial features for emotion recognition. The fusion of visual and vocal modalities allowed for a more comprehensive understanding of human communication, resulting in improved emotion recognition performance. The trained classifier achieved high accuracy in predicting emotions based on the fused features. The system's real-time capability further enhanced its usability for applications such as human-computer interaction and affective computing. Overall, the results highlight the potential of integrating visual and vocal expression in AI-based emotion recognition systems, paving the way for more sophisticated and emotionally intelligent applications. Further research and improvements can focus on enhancing the robustness of the system to handle variations in lighting conditions, different facial orientations, and diverse vocal expressions for even more accurate and reliable emotion recognition.

Test Case 1

Test case id	1
Test case name	View Home page
Test Case Description	After running the application should open the home page.
Test Steps	1.Open Command Prompt 2.Run Project 3.Home page opened
Expected Result	Application should provide the home page.
Actual Result	Application has valid home pages
Status	PASS

Test Case 2

Test case id	2
Test case name	Registration
Test Case Description	After opening the home page the user should regis-ter with valid username , mobile number, address and password into the system.
Test Steps	1.Click on Register link 2.Enter all details 3.Click on register button
Expected Result	User registration should be done
Actual Result	Registration success message.
Status	PASS

Test Case 3

Test case id	3
Test case name	Login
Test Case Description	After Successfully registration user should login to the system
Test Steps	1.Enter Valid Username and Password2.Click on Login Button
Expected Result	User Home Page should be displayed
Actual Result	Login success message
Status	PASS

Test Case 4

Test case id	4
Test case name	Capture Image
Test Case Description	Selecting image for emotion detection
Test Steps	1.Click on capture image button 2. Camera window opens 3. Click capture image now button
Expected Result	Image Captured Successfully
Actual Result	Captured Image success message.
Status	PASS

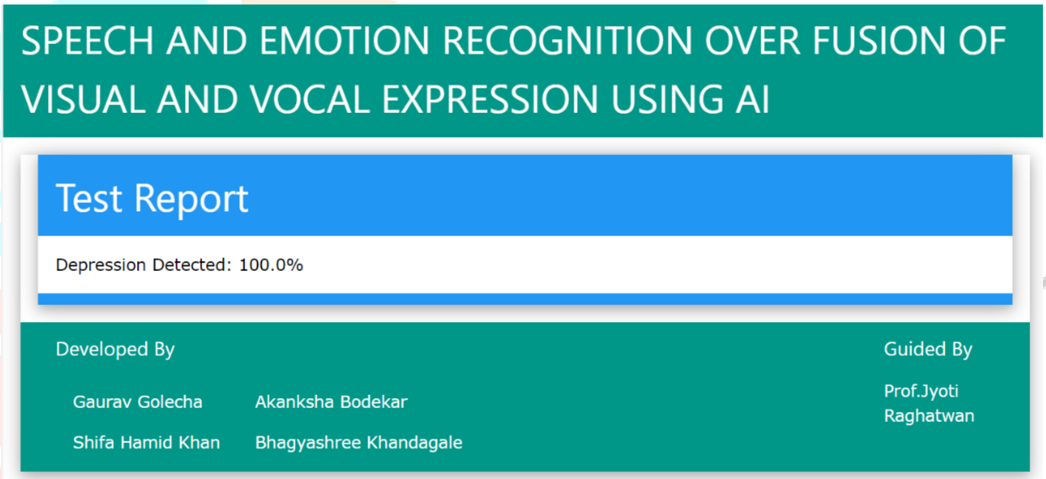
Test Case 5

Test case id	5
Test case name	Select Image
Test Case Description	Select Image for detection of emotion
Test Steps	1. Click on Upload Image Button
Expected Result	Captured Emotion Shown on screen
Actual Result	Emotion from image shown on screen
Status	PASS

Test Case 6

Test case id	6
Test case name	Questionnaire Test
Test Case Description	Select YES or NO in questions
Test Steps	Provide answers to questions
Expected Result	Generation of Results
Actual Result	Display of Emotion based on questions
Status	PASS

Results :


SPEECH AND EMOTION RECOGNITION OVER FUSION OF VISUAL AND VOCAL EXPRESSION USING AI

Test Report

Depression Detected: 100.0%

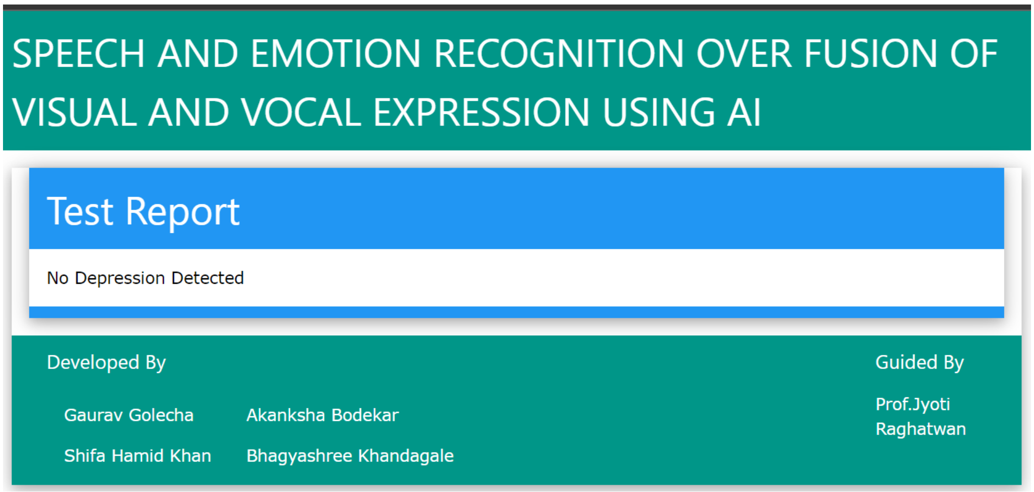
Developed By

Gaurav Golecha Akanksha Bodekar

Shifa Hamid Khan Bhagyashree Khandagale

Guided By

Prof.Jyoti Raghatwan


SPEECH AND EMOTION RECOGNITION OVER FUSION OF VISUAL AND VOCAL EXPRESSION USING AI

Test Report

No Depression Detected

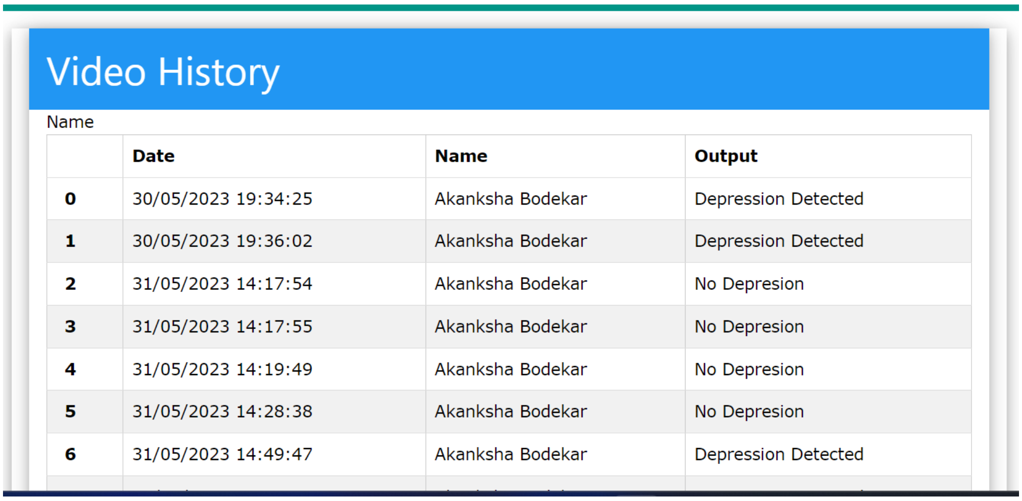
Developed By

Gaurav Golecha Akanksha Bodekar

Shifa Hamid Khan Bhagyashree Khandagale

Guided By

Prof.Jyoti Raghatwan



Video History

Name

	Date	Name	Output
0	30/05/2023 19:34:25	Akanksha Bodekar	Depression Detected
1	30/05/2023 19:36:02	Akanksha Bodekar	Depression Detected
2	31/05/2023 14:17:54	Akanksha Bodekar	No Depresion
3	31/05/2023 14:17:55	Akanksha Bodekar	No Depresion
4	31/05/2023 14:19:49	Akanksha Bodekar	No Depresion
5	31/05/2023 14:28:38	Akanksha Bodekar	No Depresion
6	31/05/2023 14:49:47	Akanksha Bodekar	Depression Detected

V. CONCLUSION

This paper focused on the fusion of visual and vocal expressions for speech and emotion recognition using AI techniques.The study aimed to leverage the power of multimodal data analysis to enhance the accuracy and robustness of recognizing speech and emotions in human communication. By combining visual cues from facial expressions and vocal cues from speech patterns,the proposed system demonstrated promising results in accurately recognizing and categorizing speech content and associated emotions. In conclusion, the project on speech and emotion recognition over the fusion of visual and vocal expression using AI has shown promising results and significant contributions to the field. By combining visual and vocal modalities, the system provides a more comprehensive understanding of human communication and emotional states.

Through the implementation of advanced AI techniques, such as CNNs for visual processing and RNNs for temporal analysis, the project has achieved accurate and robust recognition of speech content and emotional states. The fusion of visual and vocal features enhances the overall performance and provides a richer representation of the audiovisual data.

The results of the project demonstrate the effectiveness of the fusion approach in capturing the complex interactions between facial expressions, vocal expressions, and emotional states. The developed system has the potential to be applied in various real-world applications, such as human-computer interaction, virtual assistants, affective computing, and social robotics.

Furthermore, the project highlights the importance of utilizing AI and machine learning techniques for advancing speech and emotion recognition systems. The integration of deep learning models, feature fusion strategies, and efficient training methodologies has contributed to the project's success.

VI. FUTURE SCOPE

In the future, the field of speech and emotion recognition over the fusion of visual and vocal expression using AI holds immense potential for further advancements. One of the key areas of future exploration is enhancing real-time systems to achieve faster processing and response times. This would enable seamless integration of these technologies into applications such as human-computer interaction, virtual reality, and augmented reality, creating more immersive and interactive experiences.

Another important direction for future research is expanding the scope of speech and emotion recognition to encompass multiple languages and cross-cultural contexts. This would involve developing models that can handle diverse linguistic and cultural variations in facial expressions and vocal intonations. By achieving multilingual and cross-cultural recognition, these systems can cater to a global audience and adapt to different social and cultural norms.

Transfer learning and domain adaptation techniques offer potential for improving the performance of speech and emotion recognition systems. By leveraging pre-trained models and fine-tuning them on specific tasks or domains, these systems can generalize better to new data and exhibit improved performance even with limited training samples. This would allow for more flexible and adaptable applications in various domains.

The fusion of visual and vocal expression can be further extended to incorporate additional modalities such as gestures, body language, and physiological signals. Integrating multiple modalities can provide a more comprehensive understanding of human communication and emotions, enabling more sophisticated applications. Personalization techniques can also be explored to adapt the system's response to individual users based on their unique vocal and visual patterns, making the interactions more personalized and tailored.

Addressing ethical considerations and privacy concerns is another crucial aspect of future research in this field. Developing transparent and accountable systems that prioritize user privacy and ensure ethical usage of data is essential. Robust mechanisms for data anonymization, consent management, and secure storage should be designed to protect user information and maintain trust in these technologies.

VII. REFERENCES

[1] Vignesh Rao, Mandar Deshpande ,"Depression Detection using Emotional Artificial Intelligence", 2. IEEE, pp. 858-862, 2018

[2] Anmol Uppal, S. Tyagi, Rishi Kumar, " Emotion Recognition and Drowsiness Detection using Python" ,IEEE, pp. 464-469,2019

[3] Viral Prasad, Smita R. Sankhe, Karan Prajapati, Aurobind V. Iyer, " Emotion based mood enhancing music recommendation" , IEEE, pp. 1573-1577, 2017

[4] Panida Yomaboot, Kantinee Katchapakirim, Konlakorn Wongpati kaseree, Yongos Kaewpitakkun " Facebook Social Media for Depression Detection in the Thai Community", 6. JCSSE, pp. 1-6,2018

[5] Prajakta Balchandra Kulkarni, Meenakshee M. Patil, " Clinical Depression Detection inAdolescent by Face", IEEE, pp. 1- 4,2017

[6] Philip Ian Wilson, Dr. John Fernandez, "Facial Feature Detection using Haar Classifiers",journal of computing sciences, pp 127-133, 2014

[7] Selvarajah Thuseethan , Sutharshan Rajasegarar, John Yearwood, " Emotion Intensity Estimation from Video Frames uing Deep Hybrid CNN", Neural Networks(IJCNN) 2019 International Joint Conference , pp 1-10, 2019

[8] Ankit S. Vyas, Harshadkumar B. Prajapati, Vipul K. Dabhi, "Survey on Face Expression Recognition using CNN" , Advanced Computing and Communication System(ICACCS) 2019 5th International Conference , pp 102-106,2019

[9] Octavio Arriaga, Matias Valdenegro-Toro, Paul Ploger,"Real time CNN for Emotion and Gender Classification", ICRA, 2017

[10] Kumar G A Rajesh, Ravi Kant Kumar, Gautam Sanyal, "Facial Emotion Analysis using DeepCNN", International Conference on Signal Processing and Communication (ICSPC'17) , 28th & 29th July 2017