MULTIMODAL DEPRESSION DETECTION FROM FACIAL LANDMARK FEATURES USING MACHINE LEARNING

DR. M. NAVANEETHA KRISHNAN, M.E., Ph.D., Professor,

Department of Computer Science and Engineering

Mr. XAVIER S, **M.E** Student of Computer Science and Engineering

St. Joseph College of Engineering, Sriperumbudur, Chennai.

Abstract:

The massive and growing burden imposed on modern society by depression has motivated investigations into early detection through automated, scalable and non-invasive methods, including those based on speech. However, speech-based methods that capture articulatory information effectively across different recording devices and in naturalistic environments are still needed. This project presents a novel multi-level attention-based network for multi-modal depression prediction that fuses features from audio, video and text modalities while learning the intra and intermodality relevance. The multi-level attention reinforces overall learning by selecting the most influential features within each modality for the decision making. We perform exhaustive experimentation to create different regression models for audio, video and text modalities. Evaluations of both landmark duration features and landmark n-gram features on the DAIC-WOZ and SH2 datasets show that they are highly effective, either alone or fused, relative to existing approaches.

KEYWORDS - Depression, Multimodal, Facial Landmark Features

Introduction:

Depression, affecting 10-15% of the world's population, is a major mental disorder that burdens society economically and socially. Early detection and treatment are expensive and often delayed due to lack of trained clinicians and late diagnosis. To address this, researchers have been exploring technology-based methods for automatic depression detection, with video, text, and speech being the most promising and non-invasive indicators.

This project presents a novel framework that uses attention mechanisms to identify and extract features from various modalities to predict depression levels. The network uses low-level and mid-level features from audio, text, and video. The project aims to address the limitations of conventional speech-based depression detection methods due to differences in recording

conditions. It also emphasizes the importance of visual features in resolving the deep association between depression and facial emotions. Depression patients often display distorted facial expressions, prompting a growing trend among the vision community to analyze these emotions using high-end cameras in wearables and surveillance sectors.

Objectives:

This paper presents a novel multi-level attention-based network for multimodal depression prediction, combining features from audio, video, and text modalities while learning intra and inter modality relevance, due to the lack of diagnostic tests and subjectivity in depression detection.

Literature Survey:

CLUSTERING FACIAL ATTRIBUTES: NARROWING THE PATH FROM SOFT TO HARD BIOMETRICS. Authors: Andrea F. Abate, Paola Barra, Silvio Barra, Cristiano Molinari, Michele Nappi and Fabio Narducci. YEAR: 2020.

The paper presents an unsupervised clustering approach for face attributes recognition, focusing on soft biometric traits like nose, mouth, and hair. This approach uses transfer learning to group faces based on common facial features, providing a comprehensive description of each cluster and deep learning for task prediction in partially visible faces. This approach is useful in forensic scenarios with partial face photos or fingerprints.

EXPRESSION RECOGNITION METHOD BASED ON A LIGHTWEIGHT CONVOLUTIONAL NEURAL NETWORK. Authors: Guangzhe Zhao, Hanting Yang, Min Yu2. Year: 2020

This paper proposes a lightweight emotion recognition model to handle latency problems in natural conditions. It incorporates a densely connected convolution layer and model compression techniques, eliminating redundancy parameters. Multichannel input is introduced for improved learning ability. Experiments show the model outperforms other lightweight models, achieves higher accuracy, and reduces parameter number by 97 times. The FERFIN dataset is created for better label accuracy.

FACIAL EXPRESSION RECOGNITION BASED ON THE FUSION OF CNN AND SIFT FEATURES. Authors: Huibai Wang, Siyang Hou. Year: 2020

This paper proposes CNN and SIFT feature fusion algorithms for facial expression recognition. The first method uses a custom CNN network and Inception module to efficiently

extract global facial expression information. The second method uses cascade regression and SIFT features to concentrate key points on expression contributions. The fusion is classified using Softmax for improved accuracy. Tested on CK+, JAFFE, and FER2013 data sets the experimental results show that this method is an efficient method of facial expression recognition.

LOCAL LEARNING WITH DEEP AND HANDCRAFTED FEATURES FOR FACIAL EXPRESSION RECOGNITION. Authors: Mariana-Iuliana Georgescu, Radu Tudor Ionescu and Marius Popescu. Year: 2019

The researchers propose a method that combines automatic features from convolutional neural networks (CNN) and handcrafted features from the bag-of-visual-words (BOVW) model to achieve state-of-the-art results in facial expression recognition (FER). They experiment with multiple CNN architectures, pre-trained models, and training procedures. The local learning framework predicts class labels for each test image using a k-nearest neighbors' model, one-versus-all support vector machines (SVM), and a SVM classifier.

SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORK CONSIDERING VERBAL AND NONVERBAL SPEECH SOUNDS. Authors: Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen. Year: 2019.

This study explores speech emotion recognition using both verbal and nonverbal sounds in real-life conversations. An SVM-based verbal/nonverbal sound detector was developed, followed by a Prosodic Phrase auto-tagger to extract verbal/nonverbal segments. Convolutional neural networks were used to extract emotion and sound features, and an attentive LSTM-based sequence-to-sequence model was used to output an emotional sequence. The method achieved a detection accuracy of 52.00%.

System Design:

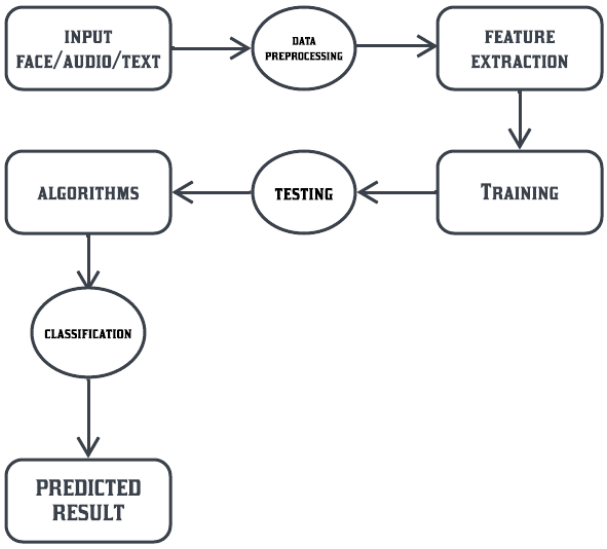
In this proposed application consists following modules

- Dataset Collection
- Data Preprocessing
- Feature Extraction
- PyAudio for Speech Depression Recognition
- Depression Recognition on Text

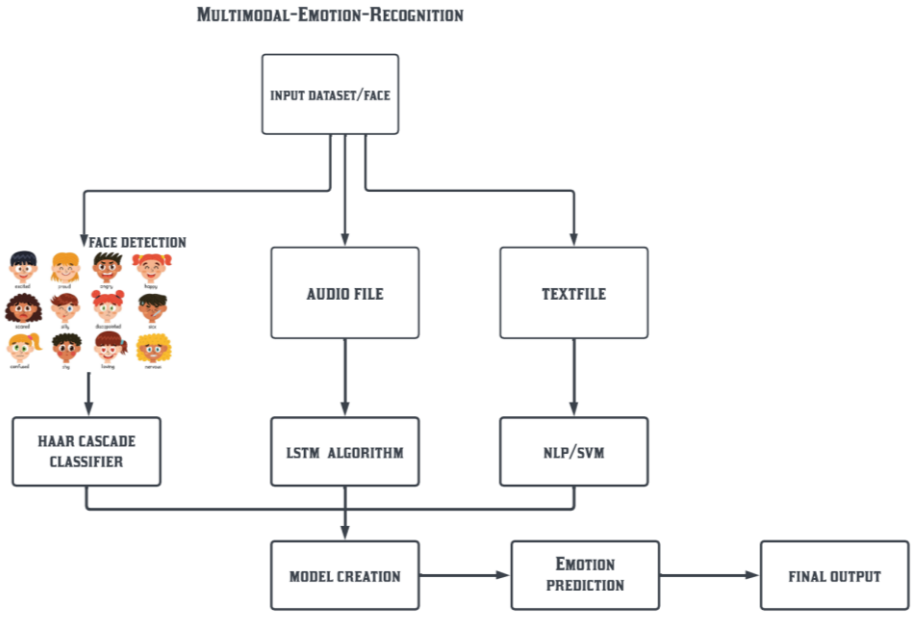
- Depression Recognition on Face

DATA FLOW DIAGRAM:

A data flow diagram (DFD) is a graphical representation of the flow of data through an information system, focusing on four main components: processes, data stores, data flows, and external entities. It is a crucial technique for modeling high-level detail, revealing relationships among and between components in a program or system. DFDs use symbols to represent these components, with processes represented by circles in DFDs.



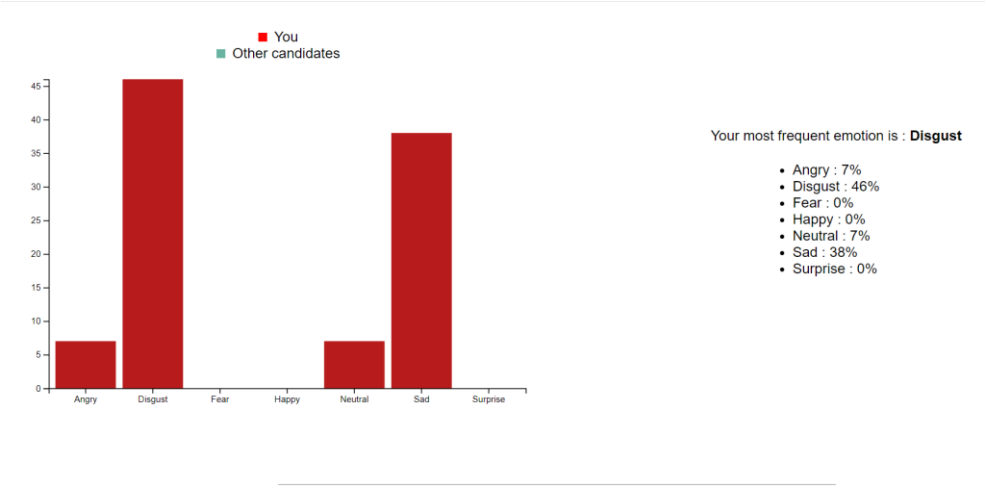
ARCHITECTURE DIAGRAM:



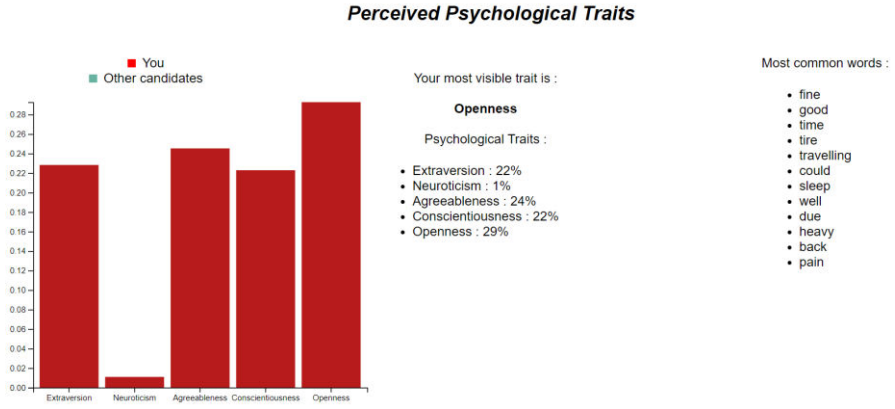
Implementation:

The first step of the system is to input an audio speech, video and text. The second phase of the system deals with quality enhancement of the input signals of the video, audio and text. The third step is Feature extraction involves the analysis of the video, speech signal and text. It is considered as an important phase of the system as extraction of relevant and significant features heavily impact on the final recognition. Some of the features extracted by various researchers are MFCC (Mel-Frequency Cepstral Coefficients), LFPC (Log Frequency Power Coefficients), pitch, energy, and voice quality. It is the main step of the system in which the video, audio speech and text is classified into different emotions based on the features extracted from the video using Haar algorithm, audio speech using LSTM algorithm and text using NLP. With the help of the features extracted, the audio speech is classified into different emotions. The extracted emotional features of video, audio and text are compared with the dataset and based on that the different emotions will be detected and result will be displayed

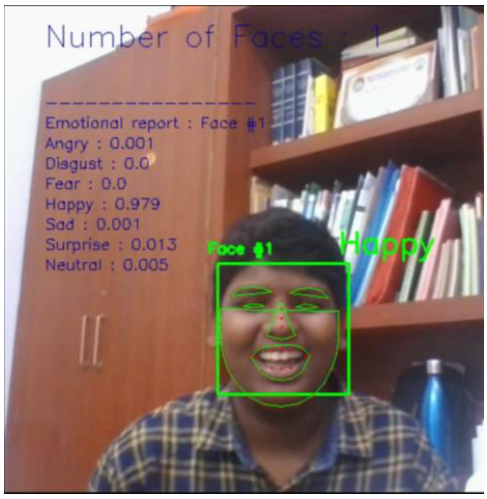
Audio - Perceived Emotion:



Text - Perceived Psychological Traits:



Video Face Detection:



Conclusion:

A multi-level attention based early fusion network which fuses audio, video and text modalities to predict severity of depression. For this task we observed that the attention network gave highest weights to the text modality and almost equal weightage to audio and video modalities. The use of multi-level attention led us to obtain significantly better results in all individual and fusion models compared to both the baseline and state-of-art. Using attention over each feature and each modality had a twofold advantage overall. Firstly, this gives us deep and better understanding of importance of each feature within a modality towards depression prediction. Secondly, attention simplified the network's overall computational complexity and reduced the training and test time.

Future Enhancement:

- We will use the system and find issues in order to improve either system architecture, algorithms for a component or library used
- To use of novel methods in one or more of our proposed system components.
- To optimize the model to achieve a greater percentage of efficiency

Author 1 :

Dr. M. NAVANEETHA KRISHNAN, M.E., Ph.D., Professor, Department of Computer Science and Engineering at St. Joseph College of Engineering, Sriperumbudur, Chennai, Tamil Nadu.

Author 2

Mr. XAVIER S, M. E student of Computer Science and Engineering at St. Joseph College of Engineering, Sriperumbudur, Chennai,

Tamil Nadu.