Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES) – An International Journal
ISSN: 2566-932X, Vol. 4, Issue 8, November 2020
Part of the Proceedings of the 1st All India Paper writing Competition on Emerging Research - PaCER 2020

Multimodal Emotion Recognition

Tejashwini N
Assistant Professor, Information Science and Engineering, Sri Krishna Institute of Technology, Bangalore, India

Kaveri A V
Information Science and Engineering, Sri Krishna Institute of Technology, Bangalore, India

Keerthana P
Information Science and Engineering, Sri Krishna Institute of Technology, Bangalore, India

Rajneesh Kumar
Information Science and Engineering, Sri Krishna Institute of Technology, Bangalore, India

Kavya C M
Information Science and Engineering, Sri Krishna Institute of Technology, Bangalore, India

Abstract: Recognizing different emotions of humans for system has been a burning issue since last decade. The association between individuals and PCs will be increasingly normal if PCs can see and react to human non-verbal correspondence, for example, feelings. Albeit a few methodologies have been proposed to perceive human feelings dependent on outward appearances or discourse or text, generally restricted work has been three models and other modalities to improve the capacities of the feeling acknowledgment framework. This paper describes the qualities and the restrictions of frameworks dependent on outward appearance or acoustic data or semantic and emotional word vector information. By the utilization of markers all over, nitty gritty facial movements were caught with movement catch, related to synchronous discourse chronicles and text inputs. The essential difficulties of feeling acknowledgment are picking the feeling acknowledgment corpora(speech database) distinguishing proof of various highlights identified with discourse and fitting decision of grouping. Feature Extraction utilized for feeling acknowledgment from video information are geometric and appearance-based while prosodic what more, phantom highlights are utilized for discourse information what more emotional and semantic word vector for text information. Later the given data is preprocessed as in called as Data Preprocessing. CNN is used to capture video and speech emotion-specific information. LSTM is used for text emotion-specific data. The basic aim of this models is to explore the capabilities of text, facial and speech features to provide emotion-specific information.

Keywords: LSTM(Long Short-Term Memory); CNN(Convolutional Neutral Network); Feature Extraction; Data Preprocessing

I. INTRODUCTION

Individuals express their own emotions through multiple modalities like human speech, facial expression, from text and body pose etc. Emotion is a strong feeling derived from one's mood. Emotion analysis of audio, visual and textual data is recognized. Linguistic analysis aims to extract the words that the user gives [1]. It aims to separate huge variation in the manner words are created chiefly in pitch commotion timing and voice quality . Between close to home human correspondence incorporates, for example, hand signals, facial communicated in language as well as non-verbal prompts. Furthermore, tone of the voice, which are utilized to communicate feeling and give input. Be that as it may, the new patterns in human PC interfaces, which have advanced from customary mouse and console to programmed discourse acknowledgment frameworks and unique interfaces intended for impeded individuals, don't exploit these significant open capacities, regularly bringing about a not exactly normal communication. In the event that PCs could perceive these passionate sources of information, they could give explicit and fitting assistance to clients in the manners that are more on top of the clients' needs and inclinations. It is accepted from psychological theory that human emotions can be classified into seven different emotions: surprise, neutral, disgust, angry, happy, scared and sad. Facial movement and the tone of the discourse play and text data plays a significant job in communicating these feelings [5-9].

Feelings are widely misused by individuals for passing on message. It is effortless undertaking for individuals however for the machine to recognize the feeling is testing. It offers a characteristics interface among machines and people, by which the framework can comprehend, decipher and react to human feelings. Emotions can considerably change the sense of the message.

From the studies we know that 7% of message is passed through spoken words, 38% is conveyed through eyes intonation and 55% facial expression. When a machine recognizes the emotion either by facial expression or by speech, it could be used to provide help to the diseased or handicapped people [2,9].

The features of the face can be transformed and the tone and the vitality in the creation of the discourse can be purposefully altered to convey various sentiments. Human beings can recognize these signals even if they are subtly displayed, by simultaneously processing information acquired by ears and eyes [9]. In view of mental investigations, which show that visual data alters the impression of discourse [9]. Understanding the human outward appearances and investigation of articulations has numerous angles, from PC examination, feeling acknowledgment, air terminal security, nonverbal correspondence and even the job of articulations in workmanship [2]. A typical supposition that will be that outward appearance at first increasingly exact.

Detecting emotional state of person by analyzing a text document written by them to appear is challenging but also it is essential commonly because of the way that the greater part of the hours of literary articulations are immediate utilizing feeling words as well as result from understanding of ideas and connection of ideas which rare portrayed in the content report. Vocal recognition has become an important research topic in signal processing, pattern recognition, artificial intelligence and so on [1]. Feature extraction is a critical step to bridge the emotional hole between speech signals and the subjective emotions. So far, an assortment of hand planned highlights has been utilized for discourse feeling acknowledgement. However, these hand designed features are usually low-level, hence they may not be discriminative enough to depict the subjective emotions.

It is needed to develop automatic feature learning algorithms to extract high-level affective feature representations for voice recognition. To define this issue, the newly-emerged deep learning techniques provide a possible solution [4]. Among them, two typical deep leaning methods are Deep Neural Networks (DNN), and Deep Convolutional Neural Networks.

Commonly non-verbal correspondence dialects and particularly outward appearance disclose to us more than words about one’s perspective. There are been a great deal of work in visual example acknowledgement to facial enthusiastic demean or just as in single processing for sound based recognition and semantic and emotional word vector detecting the text data emotions and combining these cues. We used term “feelings” to speak to an emotional state showed by means of social signs [2] [4].

II. PROPOSED APPROACH

Our Proposed Method extracts the Features of three different types of data (speech, video, text data) given by the users then apply CNN algorithm on speech data and CNN algorithm on video data and LSTM algorithm on text data to recognize the emotion ((i) female angry, female calm, female fearful, female happy, female sad, male angry, male calm, male fearful, male happy, male sad for speech; (ii) Angry, disgust, scared, happy, sad, surprised, neutral for video; (iii) Sadness, love, joy, anger, fear, surprise for text) of the users from their respective data. Figure 1 shows architecture of the proposed system.

A. Speech Emotion Recognition

At the point when we do Speech Recognition undertakings, Mel Frequency Cepstral Coefficents (MFCCs) is the best in class include since it was developed during the 1980s [8]. This shape determines what sound comes out [3] [4]. On the off chance that we can decide the shape precisely, this should give us an exact portrayal of the phoneme being delivered. The state of vocal folds shows itself within the envelope of the brief timeframe power range and therefore the work of mfccs is to exactly present this envelope.

Convolutional neural framework is a class of significant neural networks most commonly applied to analyzing visual picture [9]. Convolutional Neural Networks are very similar to ordinary Neural

Systems yet utilizes convolution instead of general framework augmentation. Convolutional is a particular sort of direct activity convolutional neural framework involves an in and out layer and multiple hidden layers. The hidden layers of convolutional neural network commonly comprise of a progression of convolutional layers that convolve with an augmentation or other dot item. Fig 2 shows the procedure of extracting emotion from speech. The CNN model developed with 7 layers and 6 conv1d layers followed by a thick layer the model will be prepared with 'm' and 'n' ages with no learning rate plan [9]. Its loss function is categorical cross entropy and the evaluation metric is accurate.

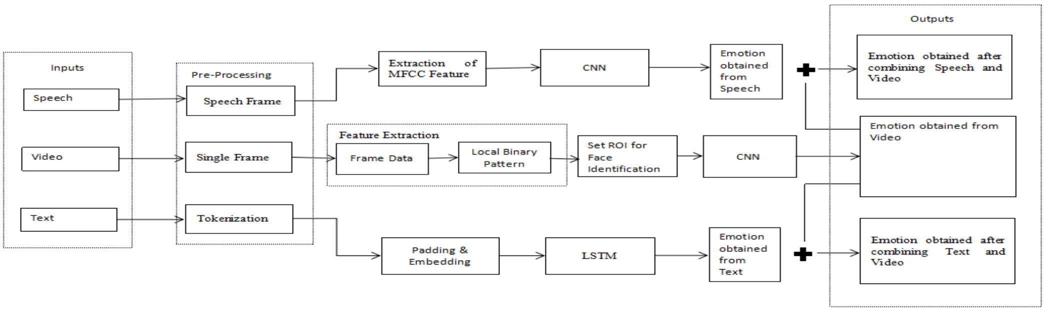

Inputs
Speech
Video
Text
Pre-Processing
Speech Frame
Single Frame
Tokenization
Extraction of MFCC Feature
Feature Extraction
Frame Data
Local Binary Pattern
Set ROI for Face Identification
Padding & Embedding
CNN
CNN
LSTM
Emotion obtained from Speech
Emotion obtained from Text
Outputs
Emotion obtained after combining Speech and Video
Emotion obtained from Video
Emotion obtained after combining Text and Video

Fig 1. Proposed System

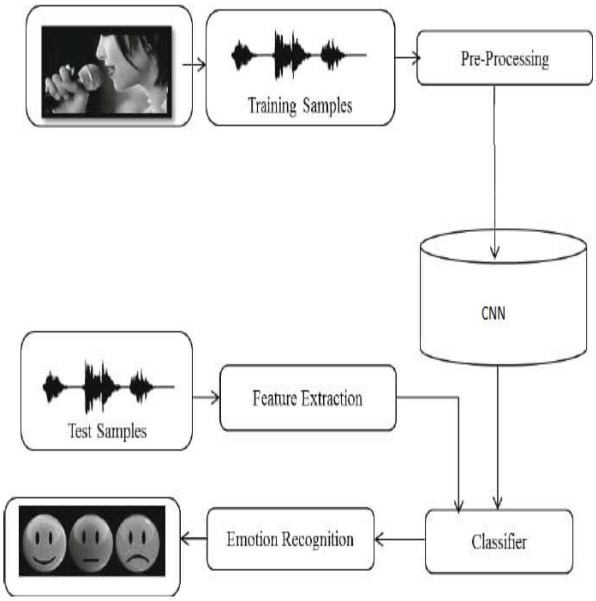


Fig 2. Framework of Speech Emotion Recognition

B. *Video Emotion Recognition*

a) *Face Identification and Tracking*

A successful article recognition strategy proposed by Paul Viola and Michael Jones in their paper which proposed object discovery utilizing Haar include based cascade classifiers. It is a approach based on machine learning where positive and negative pictures are trained by cascade function it's then accustomed to identify objects in different pictures [9].

Here we will work with face recognition. At first, the calculation needs a great deal of positive (pictures with faces) and negative (pictures without faces) pictures to prepare the classifier. By then we need expel features from it [5]. For this, Haar features showed up in Fig 3 are used. They are much the same as our convolutional kernel. Each element of face is a solitary worth acquired by taking away whole of pixels under white square shape from total of pixels under dark square shape [2] [5] [9].

Features are calculated with the help of all possible sizes and locations of each kernel. Entirety of pixels under white and dark square shapes are required for include count. To decide this they introduced the basic pictures. It streamlines count of entirety of pixels, how enormous might be the quantity of pixels, to an activity including only four pixels [2]. It boosts the speed. But among all calculated features, most of them are useless. Property that the region of the eyes is often darker than the region of the nose and cheeks is focused in first feature. Property that the eyes are darker than the extension of the nose is discussed in second feature yet similar windows applying on cheeks or some other spot is unimportant.

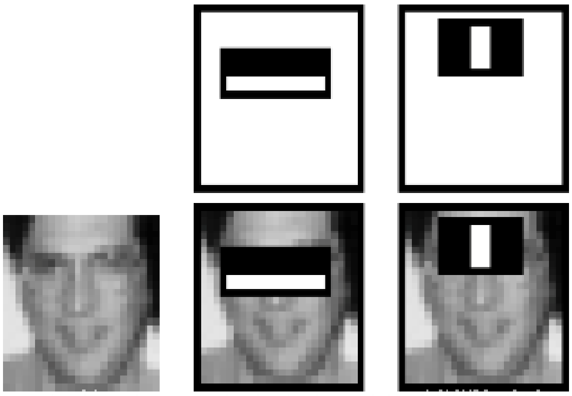


Fig 3. Features of the face

For this we apply each component on all the preparation pictures for each element it finds the least complex limit which can group the appearances to positive and negative yet clearly there will be mistakes or misclassifications. We select the features with minimum error rate, which means they are the features that best classifies the face and non-face images [5]. (The process is not as simple as this. Each image is given an equal weight in the start. After each classification, weights of misclassified images are increased [9]. Then again same process is done. New error rates are calculated. Also new weights. The procedure is proceeded until the necessary exactness or blunder rate is accomplished or required number of highlights are found).

Final classifier weighted sum of those weak classifiers. its called weak because it alone can't classify the image, but alongside others forms a robust classifier. The paper says even 200 features provides detection with 95% accuracy. Their final setup had around 6000 features. (Reduction from 160000+ features to 6000 features, i.e., big gain).

So now you're taking a picture . Take each 24x24 window. Apply 6000 features thereto . Check if it's face or not. In a picture , mostly the region is non-face region. So it's a far better idea to use an easy method to see if a window isn't a face region [5]. If it isn't, dispose it in single shot. Do not proceed again with it. Instead specialize in region where there may be a face. This way, we will find longer time to examine a possible face region.

For this they presented the idea of Cascade of Classifiers. Rather than applying all the 6000 highlights on a window, bunch the highlights into various phases of classifiers and apply individually. (Typically initial scarcely any stages will contain extremely less number of highlights). On the off chance that a window bombs the essential stage, dispose of it. We don't consider remaining highlights subsequently . If it passes, apply the second phase of highlights and proceed with the strategy [9]. The window which passes all stages is a face district.

Identifier had 6000+ highlights with 38 phases with 1, 10, 25, 25 and 50 highlights in initial five phases. On a mean , 10 highlights out of 6000+ are assessed per sub-window. So it is a natural clarification of how face tracking functions.

Since already many researchers are performing on object classification, person recognition and face detection, they'll save the model within the sort of cascades. Face identification utilizing Haar cascades is an AI based methodology where a cascade function is prepared with a gathering of information document. OpenCV as of now contains numerous pre prepared classifiers for face eyes and so forth [9]. We'll be using the face classifier specifically frontal face cascade and that we know that a video is continuous frame of images , the frames are going to be continuously analyzed by the model and face are going to be identified.

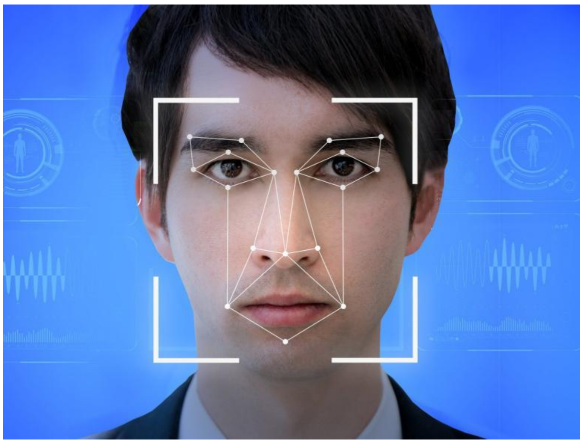


Fig 4. Region of Interest on human face

b) *Face feature extraction and classification*

This was accomplished by having inside the last convolutional layer a comparative number of highlight maps as number of classes and applying a SoftMax enactment work our proposed model for real time classification. To each diminished component map our underlying proposed design is a standard completely convolutional neural system made out of 9 convolution layers, ReLUs, cluster standardized and Global Average Pooling. This model contains approximately 600,000 parameters. It had been prepared on the IMDB sexual orientation dataset which contains 460 723 RGB pictures where each picture has a place with the classification "lady" or "man" and it accomplished an exactness of 96% in this dataset [2]. We likewise approved this model inside the FER 2013 dataset this dataset contains 35 887 grayscale pictures where each picture has a place with 1 of the resulting classes {“angry”, “disgust”, “scared”, “happy”, “sad”, “surprise”, “neutral”} [5]. Our underlying model accomplished a precision of 66% utilizing this dataset we'll allude this model as "consecutive completely CNN".

C. *Text Emotion Recognition*

a) *Feature Selection*

We will be selecting the features which influences the output of the model. We call this technique as Feature Selection. We should always only consider columns that we expect will affect the output. It will ignore Sl. No and a few other column, as emotional outcome don’t depend upon them [1].

Removing stop words sort of a , an, the….etc., we need to remove them as they could bias our model’s output. We need to consider more important and key words that we expect will have impact on our output [1].

b) *Tokenizing and Converting words to indices*

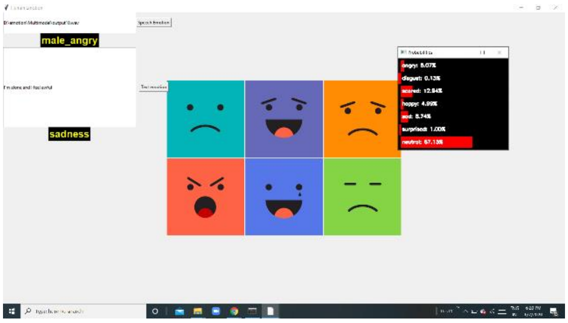
Now that, we've preprocessed words by removing unnecessary and modifying them, we now proceed and convert each word into an index. We get lists by arranging all the words in sequential order and adding +1 (index 0 unknown word). We'll provide space for each info passage with 20 words, on every given space our last unfilled information section with obscure words, possibly that we came up short on words word embedding are vectorized portrayal of words. Assume we've a space of n-dimensions, and every word in our dictionary has n dimensions and slot in our word space. This is often to preserve relative distances among words and provides a semantic understanding to our neural network. In machine learning, one-hot is a group of bits among which the legal combinations of values are only those with one high (1) bit and everyone the others low (0). Labels : { Angry, disgust, scared, happy, sad, surprise, neutral}. For instance, “anger” is 0, “scared” is 1 and “empty” is 2….etc. in an alphabetical order [7]. After this, a one-hot encoding is often applied to the integer representation. This is frequently where the whole number encoded variable is expelled, and a binary number is included for each one of a kind whole number worth.

c) *Use of LSTM*

They fall into the category called Recurrent Neural Networks. Recurrent Neural Networks will consider outlet from past timestamp as contribution for current timestamp. Due to an inner memory, which makes it totally fitted to Machine Learning issues that include successive information. Output of Embedding layer are going to be fed to the present LSTM layer [1]. It will use LSTM layer with 100 units. This layer has 100 RNN Cells, this number is variable and may be adjusted consistent with our need and complexity of our data. Input given to LSTM are going to be considered as (batch_size, timesteps, features). It's an option to modify return_sequences variable in LSTM constructor. The preprocessed data are going to be split and fed to the classifier for training then tested with remaining data.

III. RESULTS

We will be building a hybrid model that can detect the emotion of the speech, video and text input. Where the model will take both speech and video, or text and video inputs and the speech data is split into speech frames and with the help of MFCC speech features are extracted and emotions are classified. For the video input, the frames are extracted for face detection and the emotion will be recognized for the detected faces. For the text inputs, data preprocessing will be done and passed to the LSTM classifier to identify the emotion.



D:\emotion\Multimodel\output10.wav

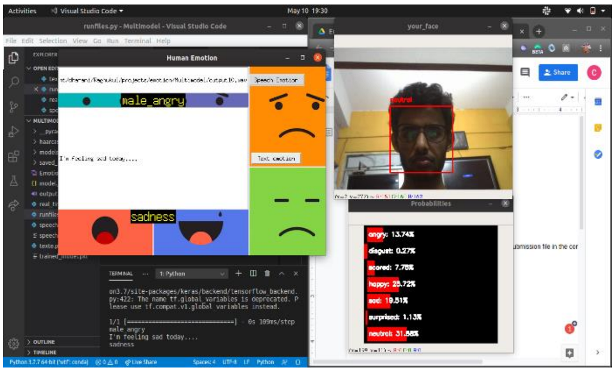
male_angry

I'm alone and I feel awful

sadness

angry: 5.07%
disgust: 0.13%
scared: 12.84%
happy: 4.99%
sad: 8.74%
surprised: 1.00%
neutral: 67.13%

Fig 5. Percentage of different Emotions obtained from speech and text



male_angry

I'm feeling sad today....

sadness

```
on3.7/site-packages/keras/backend/tensorflow_backend.
py:422: The name tf.global_variables is deprecated. P
lease use tf.compat.v1.global_variables instead.

1/1 [==============================] - 0s 109ms/step
male_angry
I'm feeling sad today....
sadness
```

angry: 13.74%
disgust: 0.27%
scared: 7.75%
happy: 25.72%
sad: 19.51%
surprised: 1.13%
neutral: 31.88%

Fig 6. Percentage of different Emotions obtained from video

IV. CONCLUSION AND FUTURE WORK

The Research on feeling acknowledgement using machine learning technique has gained the researcher's attention pretty quickly and about to become one of the powerful technologies in all fields Engineering. This paper has a brief of emotion recognition technique as to how does it work, which all emotions it can identify. The use of Machine learning in emotion recognition system is unavoidable. Here the system is taking 3 inputs speech, video and text then preprocessing the inputs into speech frames, single frame and tokenization respectively and extracts their features. After extracting, CNN and LSTM algorithms are used to classify emotions.

The models which we have implemented here are using Fastai, keras and TensorFlow and the UI part we have built using Tkinter, even though the model had performed well in detecting the emotion it is not compatible with the most of the devices like cellphones tablets etc. Hence in future using TensorFlow-lite and Android studio it can build an user friendly app which can be run on all mobile devices.

REFERENCES

[1] Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang and Qian-Bei Hong, "LSTM-based Text Emotion Recognition Using Semantic and Emotional Word Vectors", First Conference on Affective Computing and Intelligent Interaction(ACII Asia), pp.978-5386-5311-1, 2018.

[2] DuoFeng and Fuji Ren, "Dynamic Facial Expression Recognition based on two-Stream-CNN with LBP-TOP" based on 5th IEEE International Conference on Cloud Computing and Intelligence Systems(CCIS), 2018, pages 355-359, pp.978-5386-6005-8, 2018.

[3] Li Zheng, Qiao Li and Shuhua Liu, "Speech Emotion Recognition Based on Convolution Neural Network Combined with Random Forest" based on 2018 30th Chinese Control and Decision Conference(CCDC), pages 4143-4147, pp.978-1-5386-1243-9, 2018.

[4] Peng Shi, "Speech Emotion Recoginition based on Deep Belief Network" based on IEEE 15th International Conference On Network Sensing and Control(ICNSC), pp.978-1-4799-5496-4/14, 2018.

[5] Nimish Ronge, Sayali Nakashe, Asish Pawar, Sarika Bodbe, "Emotion Recognition and Reaction Prediction in videos" based on Third International Conference on Research in Computational Intelligence and Communication Networks(ICRCINN), pp.978-1-5386-1931-5/17, 2017.

[6] World Health Organization, "Mental disorders affect one in four people," Treatment Available but not Being Used., 2001. Available: http://www.who.int/whr/2001/media_centre/press_release/en/

[7] S. P. Robbins, Organizational behavior, 14/E: Pearson Education India, 2011.

[8] K. Y. Huang, C. H. Wu, M. H. Su and Y. T. Kuo, "Detecting Unipolar and Bipolar Depressive Disorders from Elicited Speech Responses Using Latent Affective Structure Model," IEEE Transactions on Affective Computing, DOI 10.1109/TAFFC.2018.2803178, 2018.

[9] T. H. Yang, C. H. Wu, K. Y. Huang, and M. H. Su , "Coupled HMM based Multimodal Fusion for Mood Disorder Detection through Elicited Audio-Visual Signals," Journal of Ambient Intelligence and Humanized Computing, Special Issue on Media Computing and Applications for Immersive Communication, vol. 8, no. 6, pp. 895-906, 2016.