International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.133.06

Analytical Comparison of Emotion using Real Time Video and Audio

Keerthana Nagula[1], K Lakshmi Priya[2], Kavya GY[3], Sunitha R S[4]
[1,2,3]*Department of Information Science and Engineering*
Ramaiah Institute of Technology, Bengaluru, Karnataka, India
[4]*Mentor, ISE, RIT, Bengaluru, Karnataka, India*

Abstract- Mood detection is an essential aspect of customer-oriented businesses. By tracking customers' preferences, businesses can cater to their customers' needs. However, data availability for sentiment analysis is limited when only textual data is used. Moreover, with an increasing number of businesses employing voice and video based services, there is an abundant source of content of customers' moods. Mood detection from video has a wide array of applications. By analyzing the emotions and expressions of customers, businesses can tailor their products and sales plans to suit specific customer needs, making market analysis more effective and cost efficient. In this project, a subject's mood is determined from real-time video, using two separate sources of information - audio and video. The results are compared and an analysis is performed to determine the most effective source for emotion detection. Two deep learning models are built, both Convolutional Neural Networks (CNNs). One model will detect the mood of a user through facial expressions captured in a real time video. The model will be trained with samples of facial expressions with different moods, with feature extraction gradient descent. The other model will be trained on voice samples and will be used to detect the mood of a user by extracting features using Mel-Frequency Cepstral Coefficients (MFCCs). The models are finally tested with standard datasets and real-time video data.

Keywords—emotion detection; CNN; analytical comparison; mood detection; deep learning

I. INTRODUCTION

Mood detection and sentiment analysis is an essential aspect of customer-oriented businesses. By tracking customers' likes and dislikes, businesses can cater to their customers' needs and preferences in a personalized manner. Most of the work done in this field identifies the mood by analyzing textual data of customers. However, data availability is limited when only textual data is used, as most customers do not provide any form of textual feedback. Moreover, with an increasing number of businesses employing voice and video based services, like voice-recognition and video interactions, there is an abundant and diverse source of voice and video content of customers' moods. Thus, classification of moods using deep learning models, using neural networks, would provide an effective and cost efficient solution to many business problems, as they require minimum human input and rapidly improve in accuracy with an increasing amount of data.

The basic concept of a neural network, in terms of the multi-class classification problem, is to train the network to make decisions in a similar manner as the human brain makes decisions. Thus, a person would identify a face emotion using features like nose, eyes, eyebrows and mouth. Similarly, the neural network is supplied a large amount of face image data, labelled with the emotion, and trained to identify the determining features and make subsequent predictions on unseen data. As such, a person would identify emotion from voice using features like pitch, voice modulation, volume, tightness and resonance. The neural network is also trained to identify such features in audio data to perform accurate prediction in a supervised manner (i.e. using labelled data). Supervised learning is preferable, as an initial step at least, to ensure that the neural network is learning the desired features (emotion), and not some other feature (like gender or ethnicity).

There are many neural network models that can perform multi-class classification. Convolutional Neural Networks, however, are the most effective neural networks for image data and most often used for audio data as well. This is primarily because image and audio data require a huge amount of information and are usually noisy and have context-related information. CNNs, which are complex feed forward networks capable of accepting large input data, effectively extract desirable features and prevent overfitting to the training data. Therefore, in this project, CNNs were chosen to build both models, after comparison with other deep learning techniques like Support Vector Machine (SVM).

II. RELATED WORK

This section discusses work done previously in this field.RenukaDeshmukh, VandanaJagtap [1] covered the use of Convoluted Neural Networks for training the model on the dataset. Support Vector Machines were used for expression classification and identifying the emotion. The dataset that was used was the Japanese Female Facial Expression (JAFEE) Database for the training and prediction.

The paper by Ajay B S, Anirudh C R, Karthik Joshi S, Keshav B N, Asha N [2] discussed how the Inception model and TensorFlow library were used to apply transfer-learning and to train the model on the Karolinska Directed Emotional Faces dataset. Pre-trained Deep Neural Networks were used.
W. N. Widanagamaachchi, A. T. Dharmaratne [3] discussed various theoretical requirements of projects in this field including Feature Extraction, Emotion Recognition and Classification by extracting the features of the face.
ByoungChulKo [4] described multiple factors that can affect the approach to this problem. It talked about using Convoluted Neural Networks with Long Short Term Memory (LSTM) and compared implemented models. It also listed out possible useful databases and included evaluation metrics to help test the built model.
Naveen Kumar H N, Dr. Jagadeesha S [5] explained the use of multimodal signals such as facial and acoustic features in recognizing human emotions. It used the Facial Expression Recognition (FER) database and discussed models built on Support Vector Machines (SVM), Hidden Markov Model (HMM) and Human Computer Interface (HCI).
Sonal P. Sumare, D. G. Bhalke [6] discussed the use of Thayer's model of mood and Hierarchical Mood Detection Framework using Gaussian Mixture Model (GMM). It used intensity, pitch, timbre and rhythm as features. These features were extracted using Support Vector Machines (SVM) with Mel Frequency Cepstral Coefficient (MFCC).
AnujaPawar [7] highlighted the use of Support Vector Machines (SVM) and Naive Bayes Classifiers for training and classifying the emotions tested. The features were extracted using the MFCC method. Bi-Directional PCA (BDPCA), Least Square Linear Discriminant Analysis (LSLDA) and Radial Basis Function (RBF) were used for classification and feature extraction.

III. MODELLING AND IMPLEMENTATION

Audio and video training and testing modules are based on Convoluted Neural Networks (CNN) with back propagation. This neural network was chosen for two reasons: CNNs work best with images and large amounts of data and since a key feature is identifying and tagging emotions, CNN is the best model for classification. A high end Nvidia GPU was used for training both models.
There are two distinct modules - The Video-based Classifier Module which detects the mood of the subject in the video clips (image frames) and the Audio-based Classifier Module which detects the mood of the subject in audio clips.

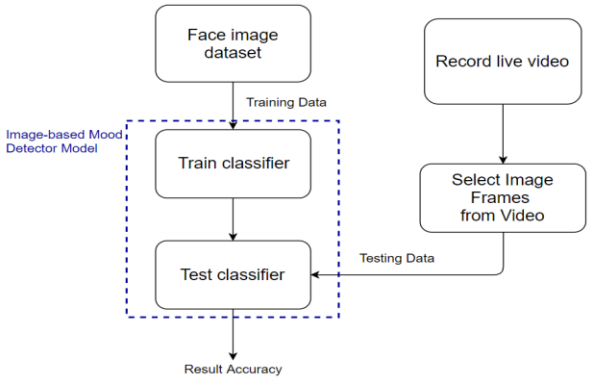

Figure 1. Video-based Mood Detection Module

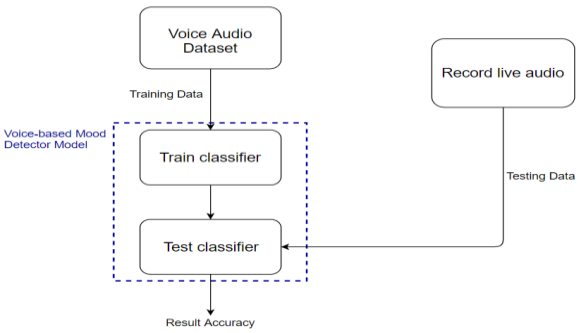

Figure 2. Audio-based Mood Detection Module

3.1 Dataset Description
Audio- Ryerson Audio-Visual Database of Emotional Speech (RAVDESS) [8]
This dataset contains 7356 files, consisting of voice samples from 24 actors (12 male and 12 female). The accents are North American and vocalize.

3.1.1 Video- FER2013 [9]
This Facial Emotion Recognition dataset consists of classified and tagged images of various emotions like happiness, sadness, anger and disgust.
Both of these datasets are standard datasets, with labeled data. Moreover, they have reasonably diverse subjects (faces and voices differ by race, ethnicity, accent and gender). They are both well balanced datasets (have proportionate number of samples for all chosen emotions).

3.2 Implementation

Input: Video clip. This video clip can be either live or recorded. The duration of the video can vary, based on the user's setting.

Output: An emotion prediction for the video and audio module each.

1. Extracting necessary data from video clip: After the video is recorded or chosen, the video and audio are separately extracted from the video clip. These two separate clips are then sent to each respective module. This done by using cv2 module for video recording and pyaudio to record audio. The video clip is split into a number of frames, and it is these set of frames that are processed by the Video Module.
2. Preprocessing and cleaning of data: Each image frame and video bit are preprocessed and cleaned, to ensure a proper training free of distortion.

Image - The images are converted into python readable data frames (arrays) using numpy. They are resized to get uniform image sizes and converted into grayscale images. Further, they are adjusted for lighting, saturation and boundary to ensure all the images are normalized before object detection. The image data is changed into categorical data frames (as we cannot pass entire sparse image arrays to CNN models). This condenses the image data into an input format compatible with CNN input layer.

Audio - These are processed to get the sample rate, so the audio can have the speed changed to normalize all the audio samples. The audio clips are transformed into python-readable data frames (using pandas).

3. Train Test Split: Both image and audio data is split into train and test split. The models made in this project were made using a train-test ratio of 8:2.
4. Feature Extraction: Video - Before building the actual emotion classification model, faces need to be identified. Thus, the first step is to extract only faces from the image. This is done using the cascade classifier, which iteratively chooses larger bounding boxes till the most accurate face is detected. It can detect multiple faces within an image. Once this is done, the features can be extracted and from the categorical features from the images. Feature extraction was done by gradient descent method.

Audio - MFCC is used to extract features from the audio clips. These features are then sent to the CNN model.

5. Building Model: Both the models are CNNs built using Keras and TensorFlow. Video - The video classifier model takes the feature data frame as its input. The model was built by adding various CNN layers (convolutional layers, activation layers, pooling layers, fully connected layer). Parameters were chosen based on the given input as well as accuracy observed.

Audio - The audio classifier model takes the extracted features of the audio and uses that to train. This model is also a CNN, with parameters chosen based on best efficiency of the model on the test data.

- Training: The models were trained with the datasets chosen. The extracted features from the training data were sent, and were divided between batches, to prevent overfitting. Training was done over multiple epochs and in an iterative method to change parameters and improve accuracy.
- Choosing parameters: Parameters like number of convolutional layers, type of activation layers and dropout rate, optimizer and loss function, differed between the models and were fine-tuned to maximize the accuracy for test data, while ensuring a good amount of generalization.
- Output Layer: The output layer for both models predicts one of the five emotions chosen - happy, sad, angry, surprised and neutral. They both use a softmax activation function, to transform the probability of the multi-class classification into a single predicted emotion label.

6. Prediction: After each model was trained, it was used to predict the class of the training data, and the accuracy was recorded by observing differences between actual and predicted value. During the fine-tuning phase, accuracies for each emotion were also recorded, to change parameters and improve accuracy for each emotion.

3.3 Results

The major result that was observed from this project is the comparison between video and audio as sources for emotion detection. The standard datasets (after adjusting to prevent overfitting) were used to test the final classifiers and the results are as follows:

Table -1 Accuracy of Models on datasets

Module	Dataset	Accuracy
Video	FER 2013	73%
Audio	RAVDESS	70%

However, the results on the real time videos used in Phase 3 of testing yielded slightly different results.

Table -2 Accuracy of Models on real-time data

Module	No of samples used	Accuracy
Video	50	78%
Audio	50	62%

These differences in accuracies between real time data and the datasets used for training could be explained by differences in broad facial structures and features for video module and accent, pronunciation and intonation (for audio module) between the subject in the datasets and the ones in real life testing. Both the results indicate that video has a better accuracy than audio to detect emotion. This difference is much starker in classifying emotions of Indian subjects. (16% difference in accuracy).

Thus, we conclude that video detection of emotion is a more accurate measure than audio detection of emotion.

The images of confusion matrix are shown below. These are plotted by calculating the percentage of misclassifications of emotions using the two classifiers, and plotting the true values versus the predicted values. The objective of the confusion plot is to visualize the patterns of classification of emotion.

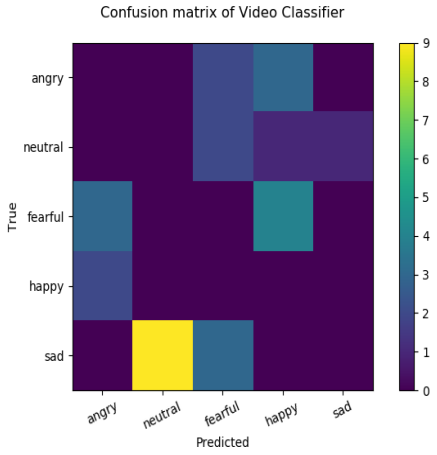

Figure 3. Confusion Matrix of Video Classifier

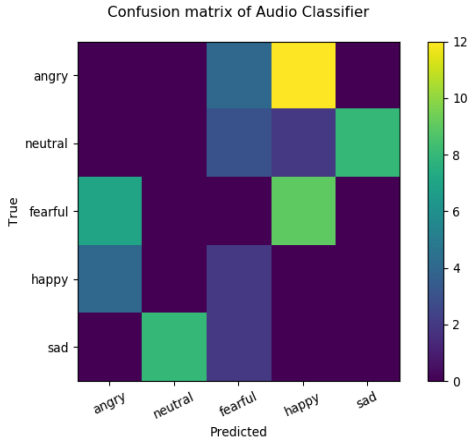

Figure 4. Confusion Matrix of Audio Classifier

3.4 Discussion

A pattern was found between detection of different emotions. This pattern was slightly different between audio and video classification. (These results are obtained from a combination of results from the standard datasets as well as real time data)

- Happiness was the most easily identified emotion using both classifiers. There was least error in identifying happiness in both face and voice, and it also detected varied expressions of happiness (i.e. different levels of happiness and different types of expressions of happiness).
- A correlation between the emotions sad and neutral was observed in both the classifiers. In video classifier, sad was very likely to be detected as neutral, while the vice versa was observed in audio classifier. Thus, sadness is a subtle emotion, and the most difficult to detect (as the default emotion is neutral).
- Anger and fear (or surprise) were more easily detected in video detection than in audio detection. There was a high rate of misclassification of anger and fear as happiness or as the other in audio model.
- Sadness was better detected using audio. The detection of sadness using video classifier was extremely difficult, usually classifying it as neutral. However, there was more accuracy in detecting sadness from the voice.

IV. CONCLUSION

This project was aimed at an analytical comparison between the effectiveness of using audio and video for detection of a person's emotion. In this project, deep learning was applied to build two (convolutional) neural network models to perform multi-class classification of emotion based on the face (video) and voice (audio) from a real-time video. Both the classifiers built were able to detect emotions from the video, with the accuracies being 73% and 70% for video and audio classifiers respectively, tested on the standard datasets.

Thus, video-based emotion detection is more accurate than audio-based emotion detection. Consequently, business applications making use of either video or a combination of video and audio as sources for emotional analysis, like customer care and consumer preferences, will get a better result than applications with only audio. Moreover, for emotions like anger and surprise, a video-based detection is comparatively more accurate in distinguishing them from each other. For detecting sadness, however, audio classification may improve accuracy.

V. REFERENCES

[1] Deshmukh, Renuka& E Scholar, M &Jagtap, Vandana. (2017). "A Comprehensive Survey on Techniques for Facial Emotion Recognition", International Journal of Computer Science and Information Security (IJCSIS). 15. 219-224.
[2] S, Ajay B. & R, Anirudh C. & Joshi, Karthik&et. al. (2017), "Emotion Detection using
[3] Machine Learning", 10.1109/ICCONS.2017.8250725.
[4] Widanagamaachchi, W. N., and Dharmaratne, A. T. (2009) "Emotion Recognition with Image Processing and Neural Networks."
[5] Ko, ByoungC.. (2018). "A brief review of facial emotion recognition based on visual information. sensors", 18(2), 401.
[6] N, Naveen Kumar & Dr. S, Jagadeesha. (2015) "Human Emotion Recognition Using Audio-Visual Modalities", International Journal of Modern Trends in Engineering and Research. ISSN: 2349-9745.
[7] Sumare, Sonal P., and Mr DG Bhalke. "Automatic Mood Detection of Music Audio Signals: An Overview", IOSR-JECE, e-ISSN: 2278-2834.
[8] Pawar, Anuja. (2017). "Recognition and Classification of Human Emotion from Audio", International Journals of Advanced Research in Computer Science and Software Engineering. ISSN 2277-128X (Volume-7, Issue-6)
[9] RAVDESS Dataset: (https://zenodo.org/record/1188976)
[10] FER2013 Dataset: (https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge)