International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 968-974

Emotion Detection Using Facial and Audio Features Using ResNet50 and CNN

Tushar Tandon, Divyansh Rastogi, Vikas Gupta, Dr. Sonu Mittal
Department of Computer Science and Engineering, Dr. Akhilesh Das Gupta Institute of Technology & Management, Delhi, India

Abstract
In this application we classify facial expressions or/and audio signals into one of the emotions (anger, disgust, fear, happy, neutral, sad, surprise). Both facial expressions and audio signals can be easily taken from user and plays vital role to determine emotion. The project has separate deep learning models to determine emotions based on facial expressions and audio signals. Both of the criteria then combined based on the probabilities of different emotions to predict final result. Emotion detection has many practical applications. This project is made using python, keras and librosa. In this project we are determining the emotions of a person in real time by using audio or video or both. This application differs from other similar application as it uses ResNet50 and CNN, we are comparing serverless to CNN with VGGFace, ResNet18, and IR50 with accuracy percentage of 65% [1], as we have not found any application that uses ResNet50 with CNN till date. We have used the technologies—ResNet50 CNN, Python, Keras and Librosa.

Keywords: Emotion Detection, Facial, Audio, CNN, ResNet50, Machine Learning, Artificial Intelligence.

Date of Submission: 02-06-2022 Date of acceptance: 14-06-2022

I. INTRODUCTION

As the exposure of machines with humans increase, the interaction also has to be-come smoother and more natural. In order to achieve this, machines have to be pro-vided with a capability that let them understand the surrounding environment. Spe-cially, the intentions of a human being. Nowadays, machines have several ways to capture their environment state trough cameras, microphone and sensors. Hence, using this information with suitable algorithms allow to generate machine perception. Emotion detection is necessary for machines to better serve their purpose. For exam-ple, the use of robots in areas such as elderly care or as porters in hospitals demand a deep understanding of the environment. Facial emotions deliver information about the subject's inner state. If a machine is able to obtain a sequence of facial images, then the use of deep learning techniques would help machines to be aware of their interlocutor's mood. It has the potential to become a key factor to build better inter-action between humans and machines, while providing machines with some kind of self-awareness about its human peers, and how to improve its communication with natural intelligence.

Emotion can be recognized easily from facial expressions or speech signals. Emo-tional displays convey considerable information about the mental state of an indi-vidual. This has opened up a new research field called automatic emotion recogni-tion, having basic goals to understand and retrieve desired emotions. Several inherent advantages make facial expression and speech signals a good source for affective computing. For example, compared to many other biological signals (e.g., electrocar-diogram), photo of faces and speech signals usually can be acquired more readily and economically. Various techniques have been developed to find the emotions such as signal processing, machine learning, neural networks, computer vision.

After this introduction section this paper has 4 more sections. In Section 2 we have written our Related Work followed by Our Approach in Section 3 and Datasets in Section 4 along with Result followed by Conclusion and Future Scope in the final Section 5.

II. RELATED WORK

Nowadays, the application of facial emotion recognition for human-computer inter-action (HCI) is becoming an emerging trend. This HCI depends to a large extent on its ability to recognize the facial expression and ability to withstand with various kinds of noise. However, confidence in its ability to provide adequate recognition remains challenging due to the variability and subtle changes of non-linear emotional features. Therefore, this paper proposed an application of using non-linear technique, Higher-Order Spectra (HOS) in recognizing the seven facial emotions (anger, disgust, fear, happiness, neutral, sadness and surprise).

2.1 Similar Work Done Before

A large amount of work has been done in the field of emotion detection using facial and audio recognition from audio and video signals written by Sai Nikhil Chennoor and his team.[2] For finding emotion using both audio and facial feature Another work related to emotion detection that we studied is Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition was written by Hengshun Zhou and his team.[1] They used face detection and alignment by Dlib toolbox, for feature extraction they used VGGFace, For feature extraction from audio they used the last pooling layer or AlexNet. They determining emotion and they find out that IR50 works better that the other two model.

However, the RNN and LSTM are shown to be limited in handling the long-term de-pendencies over the entire input sequence.[3] In this work, we propose to improve the performance of audio-visual emotion recognition using CNN and Resnet-50 models.

The primary objective is to teach a machine about human emotions, which has be-come an essential requirement in the field of social intelligence, also expedites the progress of human-machine interactions in the field of social intelligence, also expe-dites the progress of human-machine interactions. The ability of a machine to under-stand human emotion and act accordingly has been a choice of great interest in to-day's world. Recognizing emotion from speech is important aspect and with deep learning technology.[5]

Automatic sound recognition has received heightened research interest in recent years due to its many potential applications. While image classification is highly researched topic, sound identification is less mature. In this study, we take advantage of the robust machine learning techniques developed for image classification and apply them on the sound recognition problem.[6] Humans are able to comprehend infor-mation from multiple domains for e.g., speech, text and visual. With advancement of deep learning technology there has been significant improvement of speech recogni-tion.[5]

For the visual part two strategies are considered. First, facial landmarks geometric relations, i.e., distances and angles, are computed. Second, we summarize each emo-tional video into a reduced set of key-frames, which are taught to visually discrimi-nate between the emotions. In order to do so, a convolutional neural network is ap-plied to key-frames summarizing videos.[12]

Speech is a complex signal, so from the speech, we classified a signal emotion by taking some features like pitch, entropy, and energy. These speech features are lan-guage independent and non-verbal. The model resulted in an accuracy of 78.94% for speech emotion recognition on the Berlin Emo database was written by Sai Nikhil and his team.

With careful evaluation and by using the both speech-spectrogram and Log Mel-spectrogram and evaluate several facial features with different CNN models, we obtain 65.5% on the AFEW validation set and 62.48% on the test set and rank sec-ond in the challenge.[6] We experimentally show that better face recognition CNN models and choosing suitable emotion datasets to further pretrain the face CNN models is important.

III. OUR APPROACH

As one of the most successful applications of image analysis and understanding, face recognition has recently received significant attention, especially during the past several years. At least two reasons account for this trend: the first is the wide range of commercial and law enforcement applications, and the second is the availability of feasible technologies after 30 years of research. Even though current machine recognition systems have reached a certain level of maturity, their success is limited by the conditions imposed by many real applications. For example, recognition of face images acquired in an outdoor environment with changes in illumination and/or pose remains a largely unsolved problem. In other words, current systems are still far away from the capability of the human perception system. We not only categorize existing recognition techniques but also present detailed descriptions of representa-tive methods within each category. In addition, relevant topics such as psychophysi-cal studies, system evaluation, and issues of illumination and pose variation are covered.

3.1 Objective

- Determine the emotions of the person by its image of face and audio of its voice.
- In this project we are determining seven emotions that are anger, happiness, neu-tral, fear, disgust, sadness, surprise.
- Classifying an image based on its depiction can be a complicated task for ma-chines. Several human emotions can be distinguished only by subtle differences in facial patterns, with emotions like anger and disgust often expressed in very similar ways. Constraints like low latency requirement should also be considered.

3.2 Technologies Used

- **Machine learning.** It is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine

learning focuses on the development of computer programs that can access data and use it learn for themselves.

- **Deep Learning.** It is an AI function that mimics the workings of the human brain in processing data for use in detecting objects, recognizing speech, translating languages, and making decisions. Deep learning AI is able to learn without human supervision, drawing from data that is both unstructured and unlabeled.

- **Data Science.** Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

- **Neural Network.** Neural Network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. Thus a neural network is either a biological neural network, made up of real biological neu-rons, or an artificial neural network, for solving artificial intelligence (AI) problems. The connections of the biological neuron are modelled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a line-ar combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1. These artificial networks may be used for predictive modelling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.

IV. DATASETS USED

4.1 Facial Datasets

- **Kaggle Dataset.** This is the final project for DATA 622, Fall 2016 at CUNY MS Data Analytics. The data has come from a variety of sources like Labeled faces in the wild, IMFDB etc.
- **FaceDB**. The database IMPA-FACE3D was created in 2008 to assist in the research of facial animation. This dataset includes acquisitions of 38 individuals each having with a neutral face sample, samples corresponding to six universal facial expressions.
- **Jaffe.** The Japanese Female Facial Expression (JAFFE) database is a laboratory-controlled image database that contains 213 samples of posed expressions from 10 Japanese females.
- **CK+.** The Extended CohnKanade (CK+) database is the most extensively used laboratory-controlled database for evaluating FER systems.There are 593 sequences across 123 subjects which are FACS coded at the peak frame. All sequences are from the neutral face to the peak expression.
- **FERG-DB:** Facial Expression Research Group Database (FERG-DB) is a database of stylized characters with annotated facial expressions. The database contains 55767 annotated face images of six stylized characters.
- **FER2013:** The FER2013 database was introduced during the ICML 2013 Challenges in Representation Learning. FER2013 is a large-scale and unconstrained database collected automatically by the Google image search API.

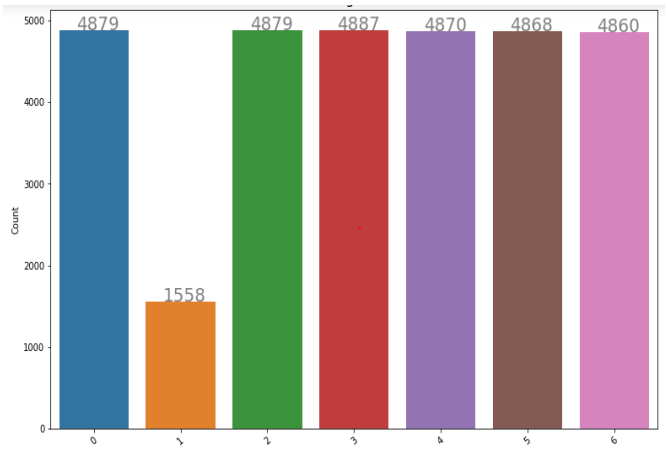

Figure 4.1: Final count of facial images per emotion
(0-anger, 1-Disgust, 2-Fear, 3-Happiness, 4-Neutral, 5-Sadness, 6-Surprise)

4.2 Audio Datasets

- **SAVEE.** Surrey Audio-Visual Expressed Emotion (SAVEE) database has been rec-orded as a pre-requisite for the development of an automatic emotion recognition system.
- **RAVDESS.** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.
- **TESS.** A set of 200 target words were spoken in the carrier phrase "Say the word _____' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions.

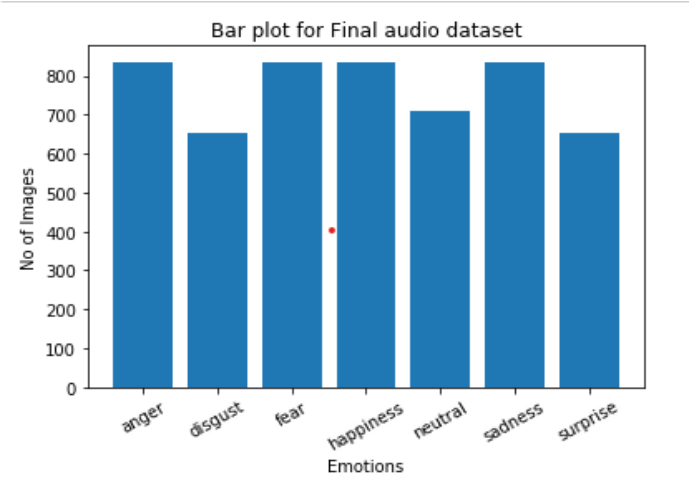

Figure 4.2: Final count of audio datasets per emotion

4.3 Collecting Datasets

For this project we have collected data from different sources for both facial images and audio clips.
For facial images we have used following datasets: Kaggle datasets, FaceDB, Jaffe, CK+, Ferg-DB and Fer2013. For audio we have used following datasets: Savee, Rav-dess and Tess. Total amount of images for facial datasets that we have used is 30,801and total amount of audio dataset is 5,356.

4.4 Preprocessing Data

- **Exploratory data analysis.** Perform exploratory data analysis on image and audio dataset to gain some knowledge about the data i.e., how data is represented, counting number of images and audio for different labels, determine whether under-sampling or oversampling is required or not etc.
- **Feature Extraction.** Variations that are irrelevant to facial expressions, such as dif-ferent backgrounds, illuminations, are fairly common in unconstrained scenarios. Therefore, before training the deep neural network to learn meaningful features, pre-processing is required. Converting images to

grey scale and extracting the face from the image using haar cascade algorithm of OpenCV library and transformations of the image dataset like resizing and scaling so that all the dataset is of same size and converting all of the images into 48p*48p. Extract features from audio like MFCC, pitch, spectral centroid and frequency by using librosa library.

- **Splitting.** Splitting data into train, test and cross validation data. We usually use 60 percent of data for training purposes and 20 percent data testing and cross validation each.

4.5 Algorithms Used

- **Haar Cascade.** Haar Cascade is a machine learning object detection algorithm proposed by Paul Viola and Michael Jones. It is a machine learning based approach, where a cascade function is trained from a lot of positive and negative images. The idea of Haar cascade is extracting features from images using a kind of 'filter', similar to the concept of the convolutional kernel. These filters are called Haar features.
- **Resnet50.** ResNet, short for Residual Networks is a classic neural network used as a backbone for many computer vision tasks. This model was the winner of ImageNet challenge in 2015. The fundamental breakthrough with ResNet was it allowed us to train extremely deep neural networks with 150+layers successfully. Prior to ResNet training very deep neural networks was difficult due to the problem of vanishing gradients.
- **Convolution Neural Network.** A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain. CNN image classifications takes an input image, process it and classify it under certain categories (Eg., Dog, Cat, Tiger, Lion). Computers sees an input image as array of pixels and it depends on the image resolution. Based on the image resolution, it will see h x w x d (h = Height, w = Width, d = Dimension) Eg., An image of 6 x 6 x 3 array of matrix of RGB (3 refers to RGB values) and an image of 4 x 4 x 1 array of matrix of grayscale image. Deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1.

4.6 Determining Performance Of Project

- **Applying ML algorithms.** Applying different deep learning algorithms, with different features. We have first train our model using only CNN and find the result with it using different parameters to find the best result. After that to increase the accuracy of our project we used Resnet50 to train our model. ResNet-50 is a convolutional neural network that is 50 layers deep. You can load a pretrained version of the network trained on more than a million images from the ImageNet database. The pretrained network can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224.
- **Determining Performance.** Determining performance of different models using various performance metrics to determine best approach. There are various metrics which we can use to evaluate the performance of ML algorithms, classification as well as regression algorithms. We have used confusion matrix to determine performance of our model. It is the easiest way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is nothing but a table with two dimensions viz. "Actual" and "Predicted" and furthermore, both the dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", "False Negatives (FN)".
- **Combining Probabilities.** Assigning weights based on importance of criteria and then give final result by combining the probabilities.

4.7 Results

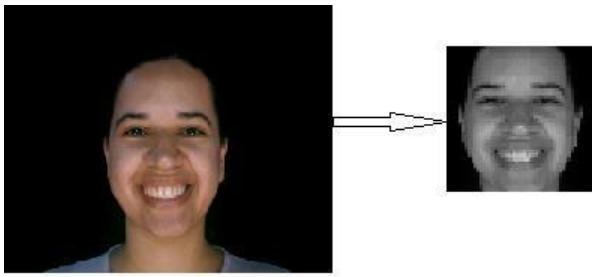

Figure 4.7.1: Preprocessing image result

Emotions using Facial Datasets:

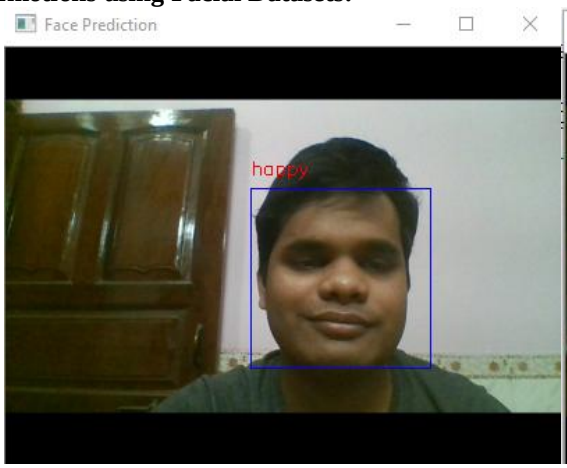

Figure 4.7.2: Emotion using facial feature: Happy

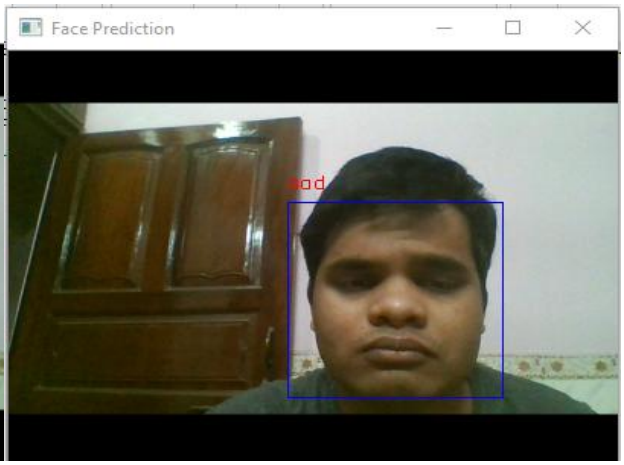

Figure 4.7.3: Emotion using facial feature: Sad

We've achieved an average accuracy of **79.02%** using facial features.

Emotions using Audio dataset: Visualizing speech signal

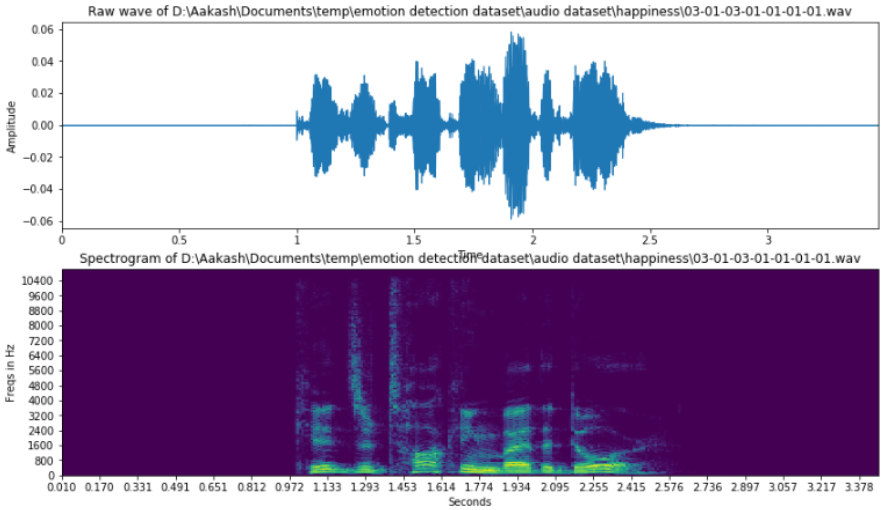

Figure 4.7.4 Amplitude and frequency spectrogram of Happiness

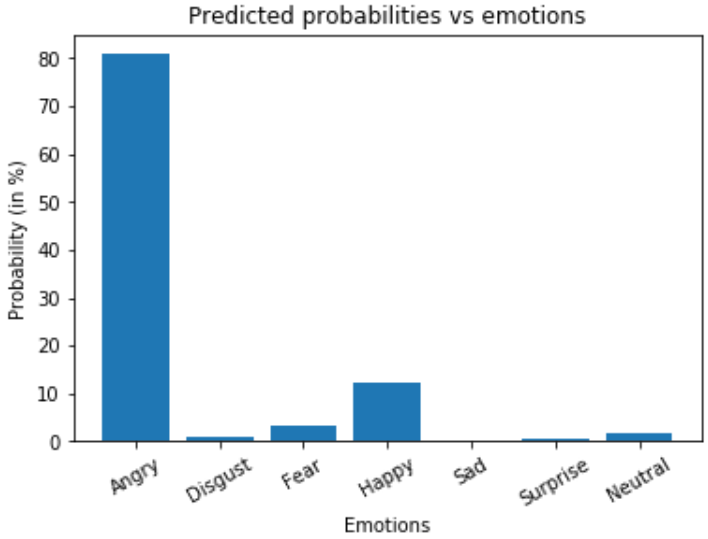

Figure 4.7.5: Emotion using audio feature: Angry

We've achieved an average accuracy of **81.12%** using audio features.

V. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

Classifying human emotions using facial expressions or/and speech signals has been researched over many past years and is still going on. It has lots of practical applica-tions in real life e.g., in market research, car safety, recommendation systems etc. Because of so many applications companies like Amazon, Google, Microsoft etc. are developing their own emotion detectors.

We have achieved an average accuracy of 79.02% on facial expression data and an average accuracy of 81.12% on speech data.

After assigning weights based on importance of criteria and after combining the probabilities, we able to obtain the emotion using both facial emotion and speech emotion.

5.2 Future Scope

Classifying an image based on its depiction can be a complicated task for machines. Since, several human emotions can be distinguished only by subtle differences in facial patterns, with emotions like anger and disgust often expressed in very similar ways and emotions such as fear is hardly detected. Therefore, we can try to over-come such factors and try to improve accuracy in future. We can work to improve our accuracy by using more datasets for both audio and image, we can choose dif-ferent techniques for feature extraction and can extract more features for audio like spectral rolloff, spectral flux, ZCR, energy etc. We can also use different deep learn-ing methods and use different machine learning techniques to increase our project accuracy.

REFERENCES

[1]. Zhou, Hengshun & Meng, Debin & Zhang, Yuanyuan & Peng, Xiaojiang & Du, Jun & Wang, Kai & Qiao, Yu. Exploring Emotion Features and Fusion Strategies for Au-dio-Video Emotion Recognition. 562-566. 10.1145/3340555.3355713. 27 DEC 2020

[2]. Chennoor, Sai & Madhur, B. & Ali, Moujiz & Kumar, T. Human Emotion Detection from Audio and Video Signals, 2020.

[3]. John, Vijay & Kawanishi, Yasutomo. Audio and Video-based Emotion Recognition using Multimodal Transformers, 2022

[4]. Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Mano-cha. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affec-tive Cues. 2020. Project Page: https://gamma.umd.edu/deepfakes/.

[5]. Mandeep Singh SCPD, Yuan Fang ICME. Emotion Recognition in Audio and Video Using Deep Neural Networks 15 Jun 2020.

[6]. Boyang Zhang, Jared Leitner, Sam Thornton, Dept. of Electrical and Computer En-gineering, University of California, San Diego. Audio Recognition using Mel Spectro-grams and Convolution Neural Netwoks, 2020.

[7]. Tarun Krishna, Ayush Rai, Shubham Bansal, Shubham Khandelwal, Shubham Gup-ta, Dushyant Goyal, Emotion Recognition using Facial and Audio features, ICMI-MLMI: Multimodal Interfaces and Machine Learning for Multimodal Interaction.

[8]. Dimmita, Nagajyothi & Siddaiah, P. Speech Recognition Using Convolutional Neu-ral Networks. International Journal of Engineering and Technology (UAE), 2018.

[9]. Bin Li, Dimas Lima, Facial expression recognition via ResNet-50, Facial expression recognition via ResNet-5, 23 February 2021.

[10]. Supreet U Sugur, Sai Prasad K, Prajwal Kulkarni, Shreyas Deshpande, Emotion Recognition using Convolution Neural Network, Department of Computer Science and Engineering (2019).

[11]. Ali, Hasimah; Hariharan, Muthusamy; Yaacob, Sazali; Adom, Abdul Hamid. Facial Emotion Recognition Based on Higher-Order Spectra Using Support Vector Ma-chines, Journal of Medical Imaging and Health Informatics. 6, November 2015.

[12]. Noroozi, Fatemeh & Marjanovic, Marina & Njeguš, Angelina & Escalera, Sergio & Anbarjafari, Gholamreza. Audio-Visual Emotion Recognition in Video Clips. IEEE Transactions on Affective Computing. PP. 1-1. 10.1109/TAFFC.2017.2713783.