CONFIDENCE LEVEL ESTIMATOR BASED ON FACIAL AND VOICE EXPRESSION RECOGNITION AND CLASSIFICATION

Mehul Naik[1], Rohan Maloor[2], Shivam Pandey[3], Dhiraj Amin[4]

[1,2,3]B.E. Student, Department of Information Technology, Pillai College of Engineering, Navi Mumbai, India - 410206
[4]Faculty, Department of Information Technology, Pillai College of Engineering, Navi Mumbai, India - 410206

Abstract - *Nowadays, a lot of Facial Expression Recognition Systems are present in the market. But those are used only for recognizing the solo expression at a time. Existing system shows only the person is happy, sad or angry etc. Hence, the existing system is programmed to recognize the Confidence of the person while talking. How confident is the person while talking? Only recognizing the specific emotion of the person is not sufficient; it could be fake by anyone. Hence all facial masks and the facial points of the person should be extracted and classified and it should be checked while interacting with the system. Along with the facial data the voice of the person provides a lot of crucial pointers that will help classify how confident the response of the person was. hence, we came up with the "Confidence Level Estimator". This system will overcome those drawbacks. By the facial expressions and voice of the person the system will display the confidence level of that person. This system can play a major role in our education system or any other competitive examination for question and answering. This will keep track of the students' expressions and will show which students are answering the questions confidently. And also, it will help to reduce the wrong practices during the tests.*

***Key Words*: Confidence level Estimator, Facial Expressions, Voice inputs, Convolutional Neural Network, Artificial Intelligence**

1. INTRODUCTION

This paper talks about the useful applications of machine learning and artificial intelligence in the field of scrutiny and security. This application is designed to be used in order to determine the credibility or legitimacy of someone's words. It makes use of both their voice as well as their facial expressions in this process. The video feed is worked on by machine learning algorithms to process individual emotions. Based on these emotions the system decides the validity of the person's claims. The use of just video would be unreliable in cases where the subject is remarkably deceptive so we also propose to record and use their voice to aid the system with the decision-making process. The video and audio are uploaded and the system processes the inputs and gives a reading that shows how confident the person is through the whole process.

2 RELATED WORKS

Estimation of Speaker's Confidence in Conversation Using Speech Information and Head Motion: Erina Kasano, Shun Muamatsu, Akihiro Matsfuji, Eri Sato-Shimokawara and Toru Yamaguchi (June 2019)

In this report, they mainly focused on the utterance and behaviour of the user's voice. They introduced a robot which can communicate to the user and give the answer to all the questions which are uttered by the user. This system is developed for the aged people who have difficulty to talk. Hence, they introduced one MMDA agent which is a robot. PRAAT is used for analysing the modulation and the pitch signals of the voice inputs. And they used the head motion for the normal head signals like detecting the nodding and the approach of the user while talking.

Emotion Recognition of Students Based on Facial Expressions in Online Education Based on the Perspective of Computer Simulation: Weiqing Wang, Kunliang Xu, Hongli Niu and Xiangrong Mia (September 2020)

In this report, CK+ and FER datasets are used which are best for any facial recognition model. CNN plays an important role in this model because deep learning is the only solution if you are dealing with any facial features problems. This system focuses on the basic seven emotions of the users. It uses the IntraFace, the publicly available package for head motion detection also for facial feature tracking for pre-processing to increase the accuracy of the model. And it also helps to process multiple faces at the same time.

Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network: Shervin Minaee, Amirali Abdolrashidi (Feb 2019)

In this report, multiple databases are used for train the model such as Ck+, FER, JAFFE and FERG which indirectly imposes the good accuracy of the model. This model gives around 99.3% accuracy which is excellent. For that they used an end-to-end deep learning framework. This framework is based on Attentional Convolutional Network.

Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion: Awais Mahamood, Shariq Hussain, Khalid Iqbal and Wail S. Elkilani

This report discusses the framework is developed with the help of dual feature fusion technique. Which initially extracts the facial landmarks with the help of the viola jones algorithm. After collecting the facial landmarks points the WLD means Weber Local Descriptor stimulation is formed.

Robust Feature Detection for Facial Expression Recognition: Spiros Ioannou, George Caridakis, Kostas Karpouzis, and Stefanos Kollias (2019)

In this report, the system emphasizes head motion and more minute features like the eye marks, mouth marks, Eyebrow marks and the nose detection. The FAPs means facial animation parameters help to find the 19 feature points to detect the more accurate emotion and expression.

3. PROPOSED SYSTEM

Our system deals with the video and audio files. Our model works in such a way that when the user uploads the video and audio of the subject, the system automatically starts processing both of them. This is followed by classifying the data into pre-existing categories. The system will later use this to estimate the confidence level. For our proposed system we have used FER2013 for the training of the face model and RAVDESS for the voice emotion recognition model. For the face model we have implemented deep learning by creating the convolutional neural network (CNN).

The face model shows the confidence of the person appearing in the video and shows the strong emotion he has while talking in the percentage value. This is done by resizing the images into the gray frame in 48X48 manner. These pixels are then converted into an array. Those numbers are then sent for prediction to our CNN model which gives the emotion of the user in the current frame. And for the confidence percentage the array of the image pixels is being converted by the formula which is derived to get predictions.

The voice model deals with the emotions of the speech it detects the emotion of the person by their speech samples. By detecting the pitches of the voice tone and short-term power spectrum of sound.

Convolutional Neural Network: It is a type of an artificial neural network used in image recognition and processing that is specifically designed for processing pixel data. It is a deep learning algorithm which takes input in the form of an image and assigns the weights and biases to various objects in the image and is able to differentiate one from the other.

MFCC: For recognizing the speech emotion in our model this feature is used. Mel Frequency Cepstral Coefficients, it originally used for the identification of monosyllabic words in continuously spoken sentences. It's derived on twisted frequency scale centred on the human auditory perception.

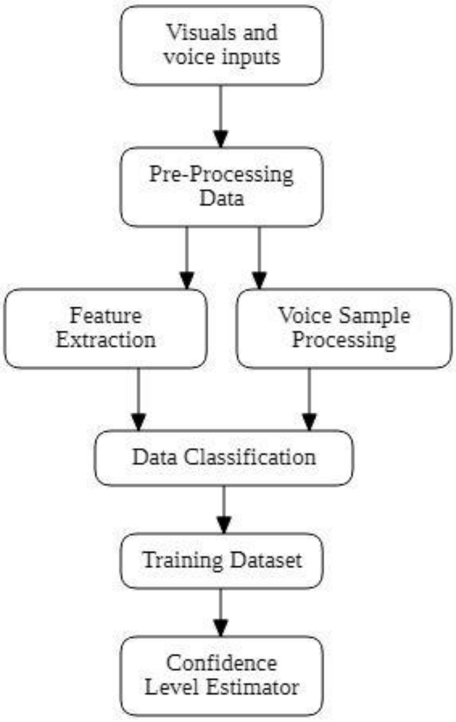


Fig. -1 Proposed system architecture

A. Visuals and Voice Input: In this process the pre-recorded video and audio files are being uploaded to the system or can be recorded and use that recorded files as an input to the voice and face models.

B. Pre-processing Data: After obtaining the facial landmarks and the face, those landmarks have to be pre-processed i.e., the irrelevant data from the input has to be removed. Like variations that are irrelevant to facial expressions, such as various backgrounds, illuminations and poses of the head, are almost same in unconstrained scenarios.

D. Feature Extraction: This process deploys the useful data which is already pre-processed. Extracting meaningful facial landmarks from the image of the face. For better results and from voice samples it extracts the different features like mfcc, chroma and mel.

E. Voice Sample Processing: This process is meant to process the incoming voice and classify the speech into different parts and focus on the pitch of the voice and the utterance and analyse the tone of the voice and recognize the emotion of the sample.

1. True Positive (TP): When actual class and predicted class data points are 1 then it is said to be true positive.

2. True Negative (TN): When actual class and predicted class data points are 0 then it is said to be true negative.

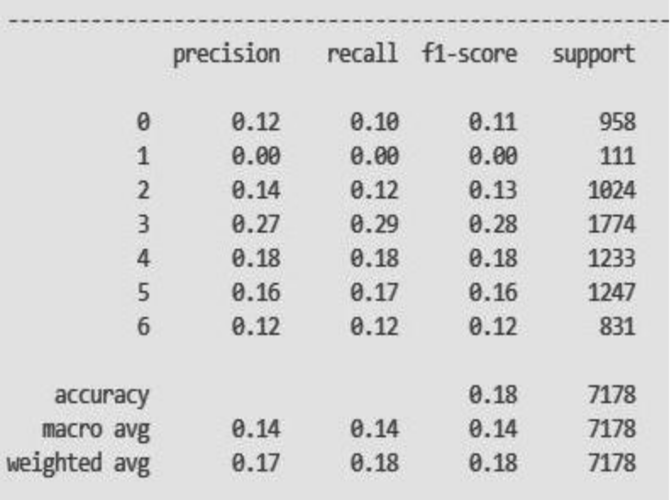
3. False Positive (FP): When actual class of data point is 0 and predicted class data points are 1 then it is said to be false positive.

4. False Negative (FN): When actual class of data point is 1 and predicted class data points are 0 then it is said to be false negative.

Precision: It is defined as the ratio of correctly predicted observations to the total predicted positive observations. The testing dataset is used to test the model and predict the accuracy.
Precision = TP/TP+FP

Recall: It is defined as the ratio of correctly predicted positive observations to the all the observations presents in that class.
Recall = TP/TP+FN

Accuracy: Accuracy means the performance measure, it defined by ratio of correctly predicted observations to the total observations.



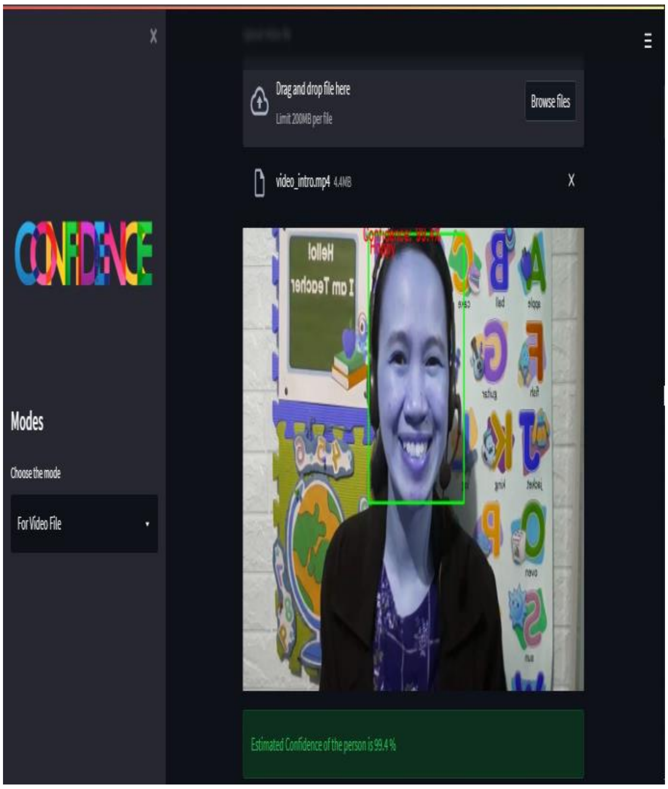
```
---------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.12      0.10      0.11       958
           1       0.00      0.00      0.00       111
           2       0.14      0.12      0.13      1024
           3       0.27      0.29      0.28      1774
           4       0.18      0.18      0.18      1233
           5       0.16      0.17      0.16      1247
           6       0.12      0.12      0.12       831

    accuracy                           0.18      7178
   macro avg       0.14      0.14      0.14      7178
weighted avg       0.17      0.18      0.18      7178
```

Fig -2: Precision, Recall, f1 score

The following snapshots describe how our models work. It takes the pre-recorded video and audio files and after and then it gives the estimated level of confidence of the person and the emotions respectively by analyzing the input.



Estimated Confidence of the person is 99.4 %

Fig -3: Working of the application

4. APPLICATIONS

This application helps to analyze the person's confidence level along with the emotions of the person while talking. This application would help to assist in the visa interview process, assist in immigration check process as well ass it can be use to assist the candidate employment interview process. By providing just a video and audio of the user it estimates the confidence level if the user in percentage.

5. FUTURE SCOPE

This application can be improved by creating our own visual dataset as well as the audio dataset for the Indian accents. Because the Indian accent is far way different than the one which we are using that is of North American accent from RAVDESS speech dataset. To make this application more dynamic it can also be equipped by making a platform where users can use this by sitting in remote places and to make it communicate both ways as a question-and-answer scenario.

6. SUMMARY

In this report, we have briefly presented an application for Identifying Confidence Level Estimators. We mentioned the objectives of the proposed system, we conducted a literature survey of previous works and at the end of chapter 2 we summarized the literature survey and listed the advantages and disadvantages of each research paper. We have briefly explained the existing system architecture and the proposed architecture. Further, the report also explains all the tools and technologies implemented in the proposed system such as Librosa, Tensorflow, Keras, Numpy and OpenCv.

ACKNOWLEDGEMENT

We would like to thank Dr. Sandeep Joshi, Principal of Pillai College of Engineering for his immense support and motivation. Our foremost thanks to Dr. Satish Kumar Varma, Head of the Department, Information and Technology of Pillai College of Engineering, for his guidance and for providing us with every facility for making and completing this project smoothly. We owe deep gratitude to our guide Prof. Dhiraj Amin. He rendered his valuable guidance with a touch of inspiration and motivation. He guided us through quite a lot of substantial hurdles by giving plenty of early ideas and which finally resulted in the present fine work.

REFERENCES

[1] Kasano, Erina, et al. "Estimation of speaker's confidence in conversation using speech information and head motion." 2019 16th International Conference on Ubiquitous Robots (UR). IEEE, 2019.

[2] Wang, Weiqing, et al. "Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation." Complexity 2020 (2020).

[3] Minaee, Shervin, Mehdi Minaei, and Amirali Abdolrashidi. "Deep-emotion: Facial expression recognition using attentional convolutional network." Sensors 21.9 (2021): 3046.

[4] Mahmood, Awais, et al. "Recognition of facial expressions under varying conditions using dual-feature fusion." Mathematical Problems in Engineering 2019 (2019).

[5] Ioannou, Spiros, et al. "Robust feature detection for facial expression recognition." EURASIP Journal on Image and Video Processing 2007 (2007): 1-22.

BIOGRAPHIES

Mehul Naik,
Student of Pillai College of Engineering, New Panvel,
Pursuing Bachelor's of Engineering Degree in IT
Engineering from University of Mumbai

Rohan Maloor,
Student of Pillai College of Engineering, New Panvel,
Pursuing Bachelors of Engineering Degree in IT Engineering from University of Mumbai

Shivam Pandey,
Student of Pillai College of Engineering, New Panvel,
Pursuing Bachelors of Engineering Degree in IT Engineering from University of Mumbai

Prof. Dhiraj Amin,
Professor at Pillai College of Engineering, New Panvel,
Department of IT Engineering