AI-Based Camera: Integrated Speech Recognition and Emotion Detection

K. Anupama Satya Sai Lakshmi, Keerthi A. M., Prasanthi V, Sai Dhruthi Varna K, K. Sudarsana Reddy
Visiting Students, Amrita University, India

***Abstract*— In the current scenario, Artificial Intelligence is emphasizing the development of intelligent machines. Considering the frequent use of web camera applications, there is a need for an enhanced and automated version of the application. Emotion Recognition also plays a crucial role in daily lives. An integrated solution of speech detection with the camera application is a great duo and handy to the user as well. In this paper, we present a method for accessing the camera to take an image and perform image processing using pre-trained voice commands. Additionally, the model also predicts the emotion of the user and produces speech output of the emotion recognized.**

I. Introduction

Recently, the use of automated devices has been increasing significantly. These days the usage of cameras is made often and there are difficulties in accessing the device directly. In this paper, our main focus is to create a user-friendly solution for accessing cameras using speech commands. This paper is a solution to joint-level speech and facial recognition for a PC with a webcam. For developing this system, the main challenge is integrating speech and face recognition as one system [1].

For Speech Recognition, the first task is to extract the features from the given speech phrase input followed by a neural network model. Mel Frequency Cepstral Coefficients (MFCC) is one of the highly efficient speech feature extraction techniques for automatic speech recognition. Its benefits include superior tolerance of noise, better capability for distinction, and simple calculation [1]

For Face Recognition, based on the speech input given to the system it performs various activities like clicking a picture followed by image enhancements or recognizing emotion [2]. If a user wants to detect emotion, then the model has to detect the face first using a Haar cascade filter. Haar Cascade classifier is widely used for applications that involve object detection. This classifier needs a large number of positive and negative images for training. It makes specific targets by examining all the features present in an image. After the face recognition, emotion is detected using a Neural network model.

The input to the proposed model is a set of 8 predefined keywords. The dataset has 34,568 one-second utterances of 7 short words- ON, ONE, TWO, THREE, FOUR, FIVE, SIX. These keywords are predicted using the keyword spotting technique. Each keyword is assigned to click a picture automatically using the webcam followed by different post-image processing techniques and emotion recognition. The keywords are assigned to automated functions like - turning on a camera and clicking a picture, RGB filter, Black &White filter, Auto-adjustment of Brightness and Contrast, Zoom In, Zoom Out, Emotion Recognition.

II. Methodology

A. Input

A large number of speech and face samples are used in this paper. In this case, around 21,264 images are used to cover all the possible facial features. The number of speech samples is 34,568 one-second utterances of 8 short words collected from thousands of people. The speech phrases are used as input to the proposed model. Fig. 1 shows the flowchart of the proposed model.

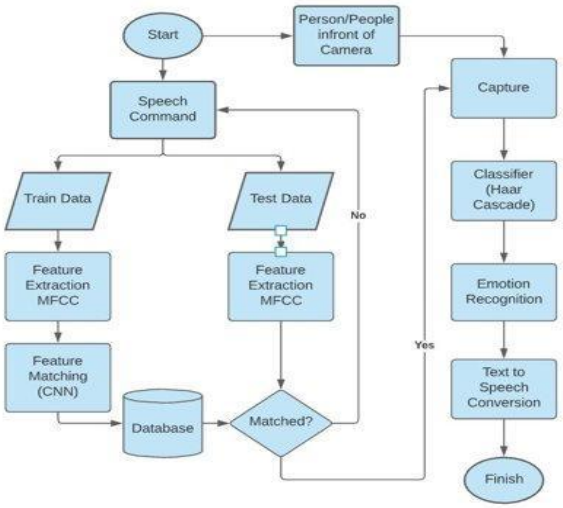

Fig. 1 Flowchart for the proposed idea.

B. Speech Recognition

A Convolutional Neural Network model is built for speech recognition using TensorFlow and Keras frameworks. Firstly, the speech data set is split into train and test data in the ratio of 80:20. The train and test data features are extracted using the Mel Frequency Cepstral Coefficients (MFCC) technique. The training data features are fed to the 3 layered convolutional neural networks and trained for 40 epochs. Later, A training speech keyword by a user is given as the input to the model. The model uses Keyword Spotting Identifier to predict the test speech input. Once the speech input is recognized, the model captures the image of the person and performs post image processing techniques and emotion recognition according to the assigned function to a keyword.

C. Face Recognition

From the flowchart in Fig.1, it can be depicted that once the keyword is recognized, this model captures the image using an integrated webcam. JavaScript is used to access the webcam. If the speech command “ON” is given, then a

picture clicked. In case if other speech commands like numbers between one to six are given then the clicked picture would be further processed. The post-processing techniques focused in this work are finding emotion and applying filters and basic camera operations to the taken image.

The functions performed by this model for the corresponding keyword inputs are listed as follows-“ONE” - RGB filter; “TWO” - Black and white filter; “THREE” - Auto-adjustment of Brightness and Contrast filter; “FOUR” - Zoom in; “FIVE” - Zoom out; “SIX” - Emotion Recognition

Emotion Recognition is done using CNN. The conventional layers used for the emotion recognition model are five. The activation function used is “Relu” and a batch size of 32. The learning rate and epochs are 0.001 and 100 respectively. The final recognized real-time emotion [2] is then produced in the form of speech. Upon detecting the emotion, the model is designed to give a speech output of the recognized emotion. In the emotion recognition model, the face of a person would be detected using Haar Cascade Classifier and that face would be given as an input to the CNN model. To train and test the CNN model we split the available data set in a 90:10 ratio. We use a large part of the data set for training and the rest for testing. This model outputs recognized emotion in the form of text. This text would be converted to speech. The final output would be a speech indicating the emotion of a person in the taken image.

III. Results and Discussions

(a) ONE-RGB filter *(b) TWO- B&W filter*

(c) THREE- Auto-adjusted filter

(d) FOUR- Zoom-in *(e) FIVE-Zoom-out*

Fig. 2. Filter output images (a) RGB filter (b) B&W filter (c) Comparison of clicked image and Auto-adjusted filter image (d) Zoom-in filter (e) Zoom out filter.

Fig.2, shows the output of different speech commands when used. The speech recognition model achieves an accuracy of 94%. Emotional recognition, the other image post-processing technique has been done using CNN model by training the model using three emotions from the Data set. Happy, sad, neutral are three emotional datasets that are used for training the model. Fig. 3 shows the percentage of emotion indicated by the model. It can be inferred that the input image to the model being happy and the model predicted it perfectly with 100% index respectively. Table 1 shows the numerical analysis of the emotion recognition model.

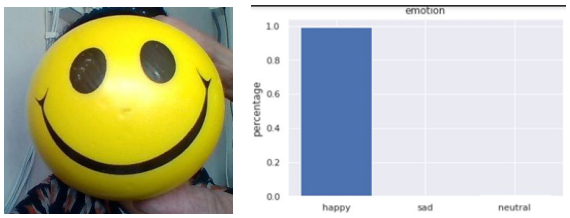

Fig. 3. Percentage of the emotion of the image.

TABLE I. ANALYSIS OF EMOTION RECOGNITION MODEL

Emotion	***Training samples total***	***Testing samples total***	***Number of correctly classified images***	***Sensitivity***	***Accuracy***
Happy	8090	899	824	$\frac{824}{899} = 0.92$	$\frac{1670}{2127} = 0.785$
Sad	5469	608	400	$\frac{400}{608} = 0.66$	
Neutral	5578	620	446	$\frac{446}{620} = 0.72$	

IV. Conclusion & Future Work

A. Conclusion

Concluding, we have developed a camera application that operates on speech commands. Automatic speech recognition helps to click images and apply five post-processing image techniques and emotion recognition. The model recognizes three different emotions namely happy, sad and neutral with an accuracy of 78.7 %. This model can also be used in a variety of real-time applications.

B. Future Work

To increase the accuracy of the emotion detection model by improving the sensitivity of sad and neutral emotions in the model by extracting special features.

V. Acknowledgement

The authors would like to thank Dr Mansour Tahernezhadi from the department of electrical engineering for his technical and moral support throughout the project.

VI. References

[1] H. Alshamsi, V. Kepuska, H. Alshamsi and H. Meng, "Automated Facial Expression and Speech Emotion Recognition App Development on Smart Phones using Cloud Computing," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2018, pp. 730-738, DOI: 10.1109/IEMCON.2018.8614831.

[2] B. T. Nguyen, M. H. Trinh, T. V. Phan and H. D. Nguyen, "An efficient real-time emotion detection using camera and facial landmarks," 2017 Seventh International Conference on Information Science and Technology (ICIST), Da Nang, Vietnam, 2017, pp. 251-255, DOI: 10.1109/ICIST.2017.7926765.