Facial Emotion Classifier using Convolutional Neural Networks for Reaction Review

Makarand Madhavi[1, *], *Isha* Gujar[2], *Viraj* Jadhao[3], and *Reshma* Gulwani[4]

[1]Department of Information Technology, Ramrao Adik Institute of Technology
[2]Department of Information Technology, Ramrao Adik Institute of Technology
[3]Department of Information Technology, Ramrao Adik Institute of Technology
[4]Department of Information Technology, Ramrao Adik Institute of Technology

Abstract. Applications of facial emotion classification is gaining popularity in the world. There are many ways to train a model to classify human facial expressions by use of existing technologies. The strategy to order and recognize feelings of an individual conveyed by his facial expression is done by contrasting it to a gathered set of labelled experiences of feelings. In this paper, we propose the making of an intelligent system that will recognize and classify facial emotions. A multi-layer Convolutional Neural Network model is proposed. Another method of training using pretrained ResNet50 Model is explored. A basic live video streaming application is developed to showcase the use case of our model which will be capable of monitoring and recording facial emotions in real time from a live video stream and subsequently summarize the overall reactions at the end of the stream.

1 Introduction

By definition, A facial expression is one or more developments or positions of the muscles under the surface of the face. Facial expressions are form of nonverbal communication.[1] These muscles create a specific visual structure on the face which relates to the specific emotion. Humans recognize an emotion by identifying the structure of the facial expression and relating to an emotion by past experiences. The way humans recognize emotions comes very naturally by the way their brains are coded. In order for machines to be able to recognize and classify emotions it needs to replicate neuron structures of the brain in a similar fashion. Convolutions Neural Networks is a technology that is widely used in artificial intelligence in image classification. It can be used to train a model to classify human emotions based on labelled set of data of images of facial expressions.

Such a system that can identify and classify a facial emotion has a wide range of applications. It can be used is psychological studies and surveys with huge sample size. An automatic emotion recognition system will efficiently be able to summarize survey data with very less time requirement. It can be used in patient monitoring systems. An intelligent emotion recognition system has a high usage in advertising and customer feedback. In movie review analysis consisting of a large number of reviewers and film duration in hours, an emotion classification system will be able to summarize audience reaction in a matter of minutes after a reviewer finish watching. A live webcam image can be obtained to quantify real time emotions which additionally have various applications.

2 Background

The most common applications of artificial intelligent systems are mimicking human behaviours in one form or another.[2] To be able to mimic such behaviours a computer needs to understand the behaviours themselves, how they look, how they act, how they feel. The very soul of why humans act the way they do is driven more by emotions than by logic. Decades of artificial intelligence research fuelled by the development and availability of high amounts of compute power enabled by GPUs have made it possible to make artificially intelligent systems capable of classifying images to classes. Such systems can be used to recognize emotions from facial images. If computers can keep track of psychological state of a user they can better serve and act accordingly.[3] Emotion recognition is therefore vital to human computer interaction.

3 Literature Review

3.1. Use of Neural Network technologies

[*] Corresponding author: makarandmadhavi99@gmail.com

Neural Networks can be used to classify facial expressions like happy, sad, angry, surprise etc. using MATLAB, and Neural Network Toolbox technologies as per research conducted by Dilbag Singh. They are however difficult to implement with complex high-density code and sub optimal processing.[4] Research done by Nithya Roopa.S1 elaborates the use of open-source technologies such as Inception Net, Tensor Flow, Kaggle with the help of Karolinska directed emotional faces (KDEF) Datasets and transfer learning algorithms, however they only achieve an accuracy of 35%.

3.2 Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews

Movie reviews can be also analysed from IMDb reviews however this research argues that the reviews are highly user taste specific containing user biases with halo and horn effects.[5] A more direct reaction capturing system will be more accurate to get un-biased and organic reviews.

3.3 Reaction Review

Movie test screening involves three phases:

1. In Before Phase: The audience is informed about the movies and is advised about a few things.
2. In Watching Phase: The facial emotion recognition system views and records the facial emotional reactions of the audience. The audience watches the movie and their emotional reactions are recorded and saved. The system takes reviews from the audience or provides feedback. The audience is given a questionnaire to review the film which gives clarity of the plot and other general information.
3. In After Phase: After collecting feedback, the recorded emotional reactions and reviews from the audience are analysed to prepare a report.

3.4 Supporting Technologies

A web-based user interface system will be most suitable for a facial emotion recognition system as it can be further scaled up for specific applications as per the need.[6] Python programming language is widely used in data science and artificial intelligence; we will be using python library Flask supported by other open-source libraries to enable us to train and use a facial emotion recognition model. Having a central backend for the highly computational emotion recognition module will be efficient in order to scale up in future scope.

4 Methodology

First, we are preparing dataset for training and testing models. We will prepare two models and pick the best among the two. The first is the CNN model and the following is transfer learning model with a pre-prepared Resnet50 model.[7] Then we implement a module for prediction emotions from images utilizing trained models which are used for recognizing emotions from frames of video and then visualizing predicted emotional data.

4.1 Datasets

Neural networks require large amounts of data for training and validation. The selection of high labelled data is directly responsible for the performance of the model. Thus, we require high quality and quantity of data for training and validation. There are quite a few datasets available for research like the KDEF dataset and FERC-2013 dataset. The Facial Expression Recognition Challenge (FERC-2013) dataset is more suitable for this task. This dataset contains cleanly labelled 32000 in number 48x48 resolution images. [8,9,10] The FERC-2013 set shows emotions from all genders, races and age groups. This dataset might be difficult to interpret however the diversity of the data will be beneficial for our model. We thus hope to achieve good results from this dataset. FERC-2013 dataset is employed for training the model, the dataset contains 48x48 grayscale images categorized into these emotions – happy, sad, angry, neutral, fearful, surprised and disgust. We will be training a neural network model using deep learning with the assistance of the TensorFlow library. Using Keras module that uses deep learning to train a neural network model.

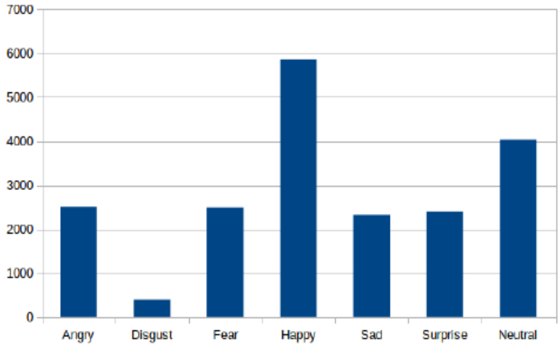


Fig. 1. FERC2013 dataset summary.

4.2 Convolutional Neural Network Model

We are going to construct a CNN model with the following architecture. Having four convoluted layers, a dense layer and a final seven output SoftMax layer denoting the expressions - angry, happy, sad, disgust, fear, surprise and neutral. [11,12] The filters range from 2x2 to 3x3 filters with 64-128 filters on each convolutional layer. We also add dropout and max-pooling to improve training efficiency and avoid overfitting of the model. The images before being fed to the model are pre-processed by rescaling them to a factor of 1.0/255 for each pixel value so that they are in 0-1 range. This is done since we are training the model from scratch and it is generally good for computation.[13] This also eliminates bias for darker and lighter images such that now every image had equal impact on the model. If images are not rescaled, the brighter images have more impact on the model than lighter images. This also eliminates racial biases in training. So, a pale complexion of skin colour is given the same weightage as a darker complexion of skin colour. We also implement some image augmentation techniques to prevent overfitting the model and ensure features learned are not highly positional but relative in nature. Image augmentation also virtually increases the number of images in the dataset as each augmented image is treated as a different image. We randomize width shifting, height shifting and zooming by a factor of 10%. We also randomly flip images horizontally. The loss function used is categorical cross entropy and optimizer for training used was Adam optimizer with a relatively slow learning rate. [14,15] The model was tried to train with Stochastic gradient descent algorithm but the training turned out to be too slow with a lower learning rate, and unstable with a higher learning rate. The training and validation loss and accuracy curves followed each other for a few epochs after which the validation loss was seen to be incrementally increasing which denoted that the model is overfitting beyond this point and couldn't get any better. We are achieving 66.23% accuracy with this model.

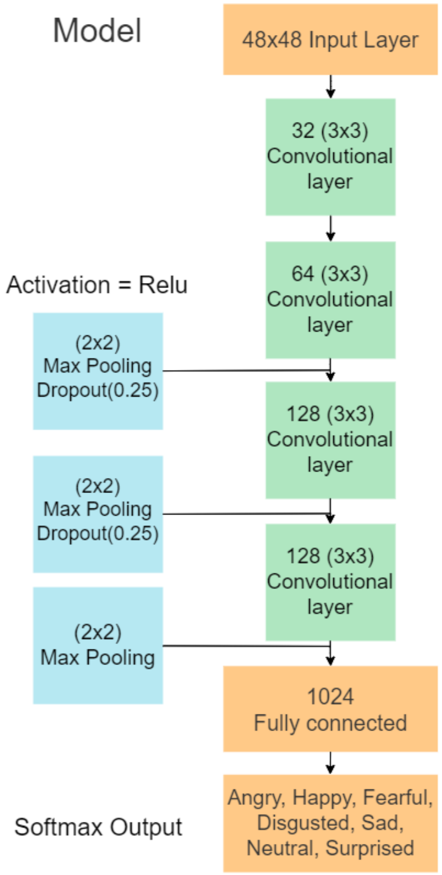

Model
48x48 Input Layer
32 (3x3) Convolutional layer
Activation = Relu
64 (3x3) Convolutional layer
(2x2) Max Pooling Dropout(0.25)
128 (3x3) Convolutional layer
(2x2) Max Pooling Dropout(0.25)
128 (3x3) Convolutional layer
(2x2) Max Pooling
1024 Fully connected
Softmax Output
Angry, Happy, Fearful, Disgusted, Sad, Neutral, Surprised

Fig. 2. Structure of CNN model used for training.

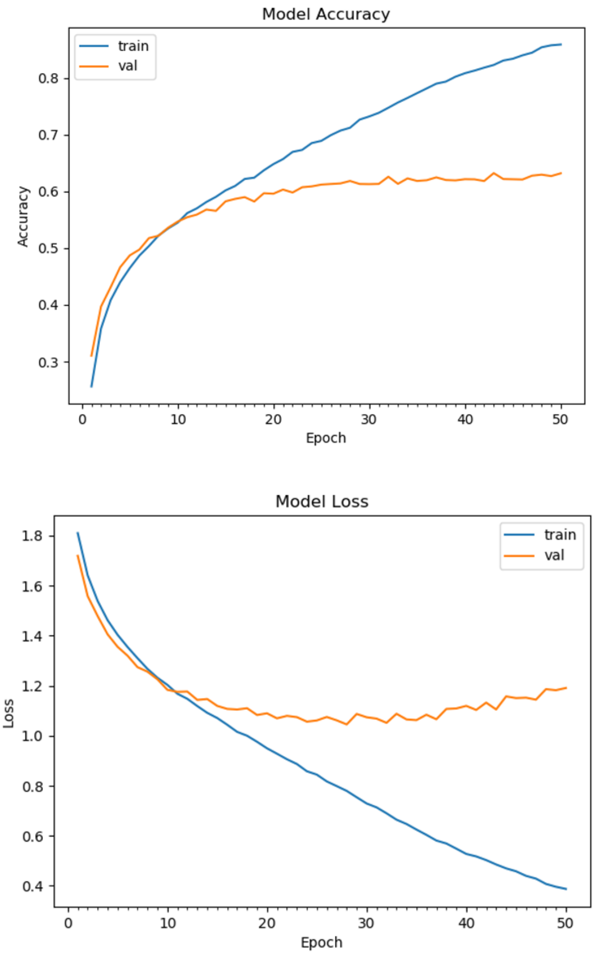


Fig. 3. Accuracy and loss curve after training.

4.3 ResNet-50 Model

ResNet, which stands for residual networks, is widely used CNN model for computer vision tasks. This model was awarded the ImageNet Challenge in 2015.[16] This model has 50 convolutional layers. ResNet solved the problem of vanishing gradients which resulted for easier training. We are using the ResNet50 variation of this model for our application. We use pretrained "ImageNet" weights loaded into the model before our training and fine tuning. We add 2 more dense layers on top of this model. The first dense layer is equipped with a L2 class layer weight regularizer. This is added to apply penalties and regularize layer output. Lastly, we add a 7 neuron SoftMax activation output layer representing our 7 emotion classes. Dropout layers with a factor of 0.2 are added in-between the dense layers as well to prevent overfitting. All the layers in this model are set to non-trainable except for last four layers for initial training. [17,18] Same image augmentation techniques are used on the images that was discussed in the previous model consisting of rescaling, width shifting, height shifting, zooming and horizontal flip. The ResNet50 model is designed to take RGB images as input. It takes 3 layers of RGB as input but our dataset images are grayscale having only one grayscale layer.[19] To overcome this issue, we send the same grayscale layer to all 3 RGB input layers. Once again, the Adam optimizer seems to be working the best here. The loss function is categorical cross entropy. We train for 50 epochs. The accuracy gradually increases and loss gradually decreases for a few epochs after which we don't see any improvement in training.

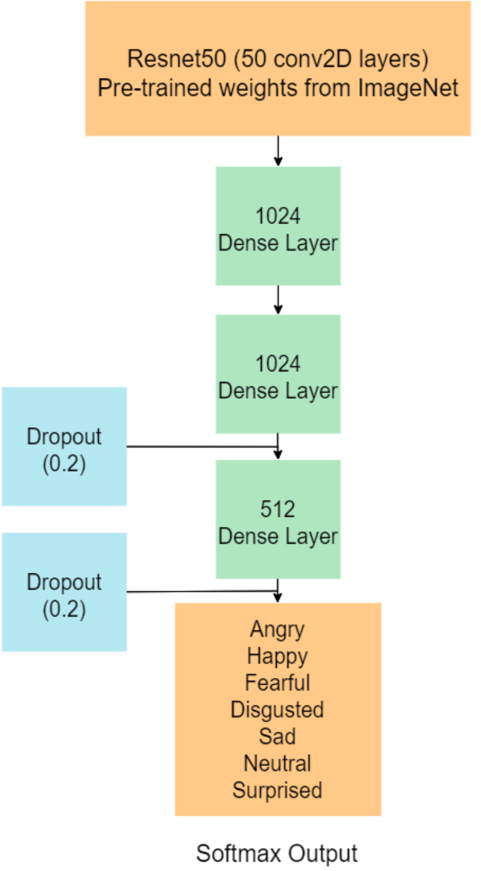
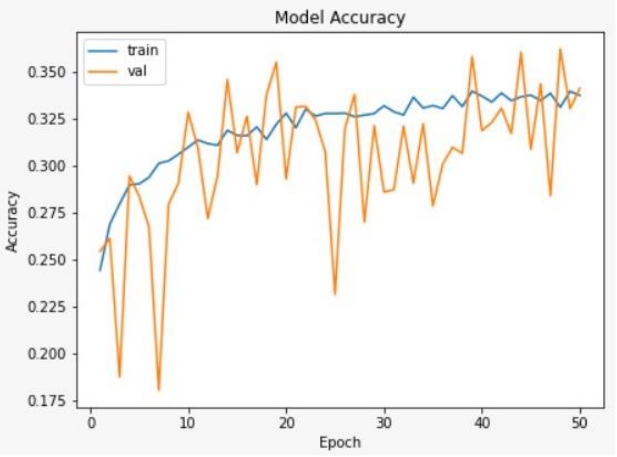


Fig. 4. Resnet50 model by Transfer Learning.



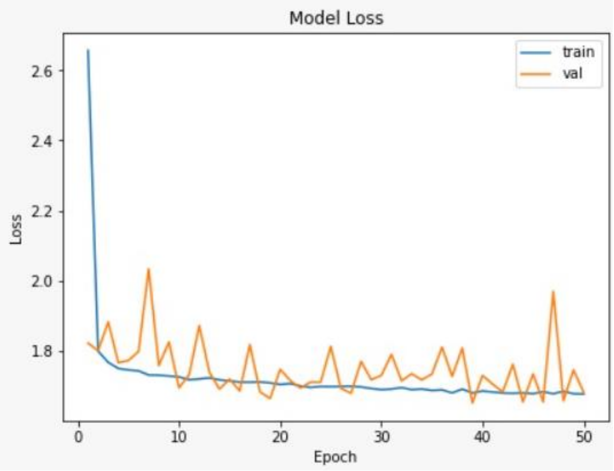


Fig. 5. Accuracy and loss curve after first training ResNet.

We get 33.25% accuracy from this training. We then try to fine tune this model by setting all layers of the model to trainable. Using Adam optimizer and a very low learning rate we train for another 50 epochs.

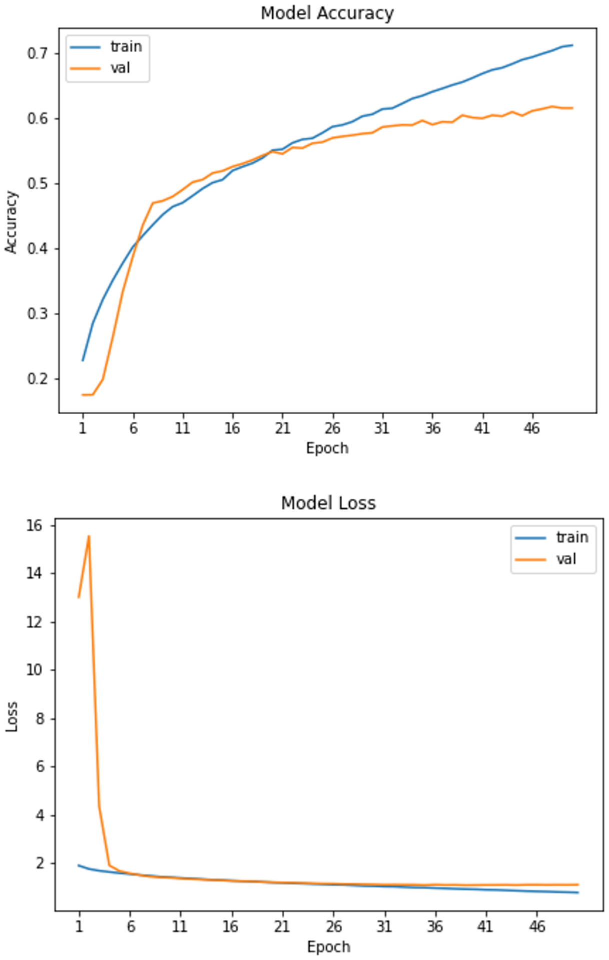


Fig. 6. Accuracy and loss curve after fine tuning ResNet50 model.

The accuracy gradually increases and loss gradually decreases for both training and validation. The training and validation curves follow each other for a few epochs after which training curves continue to steadily improve with no improvement in validation curves which denotes the model is overfitting beyond this point.[20] We achieve an optimum validation accuracy of 61.47% after fine tuning.

4.4 Predicting Emotions

For prediction, input images are converted to grayscale since our model is trained on grayscale images. All the faces inside an image frame are detected using the OpenCV Haar cascade classifier. Detected faces are scaled to 48x48 resolution images which is the same resolution of images our model was trained on. Scaled images are run through the model to classify emotions.

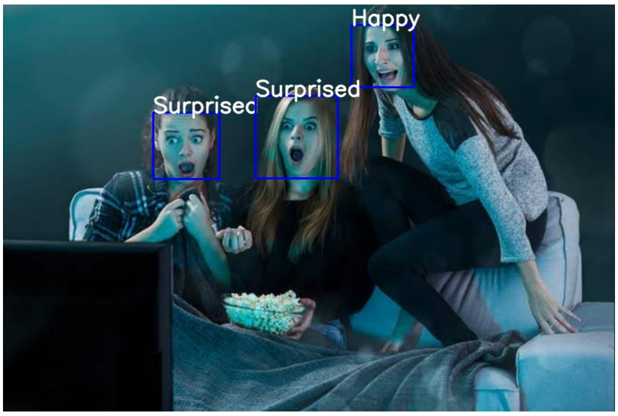


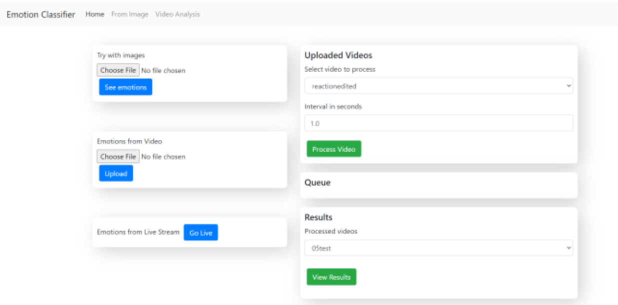
Fig. 7. Faces and emotions classified by CNN model.

5 Application

For detecting faces and classifying emotions from a video stream we sample the video into image frames at regular intervals. These sampled images are run through a processing pipeline that classifies emotions as described above. Sampling every frame of the video stream and processing it to detect faces and classify emotions requires high memory and compute power. Given that in video consecutive frames are very similar in nature [21], processing a similar frame which has a high probability of giving similar results in use cases where precision isn't important is a waste of resources. Thus, we can sample image frames at an interval of 0.5 seconds or 1 second depending on the use case. For e.g., movies that have a runtime of over 120 minutes can choose to sample audience reactions at every 2 or 5 seconds.

The web-based GUI is created using flask at the backend for handling requests and connecting with our modules for the prediction of emotions from input images, videos and livestreams. [22] Using OpenCV library we were able to take input videos and images stored in any format. Web based GUI's however can only display images of PNG, jpeg or SVG format and videos with X264, 3GP or MPEG4 format with MP4, WebM, or Ogg containers. We therefore create predicted and processed video in mp4/H.264 format using opensource openh264 from windows utility; predicted images in PNG format.

We are also able to predict emotions live on the go and record them from user's video capture device or screen share device. This is made possible by capturing images at regular intervals from the capture device. The user is given the option to set the interval at which frames are sampled. Sampled images are sent to the prediction server via HTTP post requests. [23,24] The images are processed by the emotion recognition system in a similar fashion as stated in prediction. The images are stacked in a mp4/H.26 container to generate the output.



Try with images
No file chosen

Emotions from Video
No file chosen

Emotions from Live Stream

Uploaded Videos
Select video to process: reactionedited
Interval in seconds: 1.0

Results
Processed videos: 05test

Fig. 8. Main Page.

5.1 Application Results

The videos are passed frame by frame into the model to predict emotions. The output format we get from video processing module is as follows: -

Table 1. Output from video processing module.

Timestamp	0.5	1
Happy	1	0
Disgusted	0	0
Fearful	0	0
Neutral	0	0
Sad	0	0
Surprised	0	1
Angry	0	0

Emotion probability is plotted against time for each emotion for analysis. Prediction data is aggregated in a csv file for recording and further analysis as required.

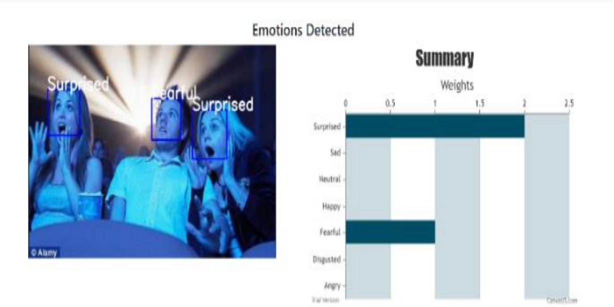


Fig. 9. Emotions detected in image.

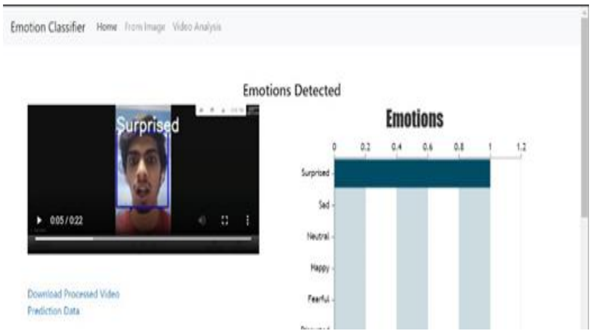
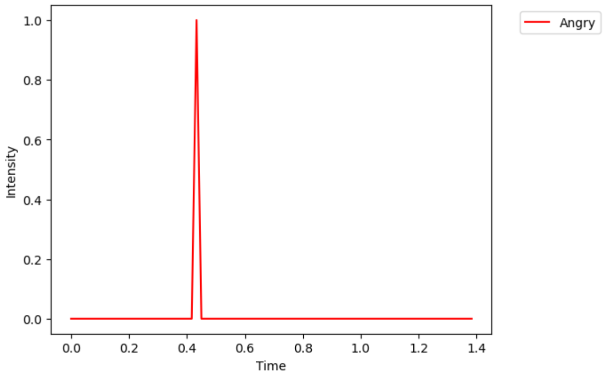
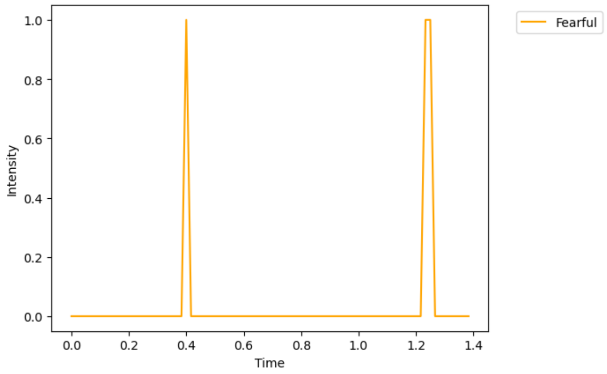
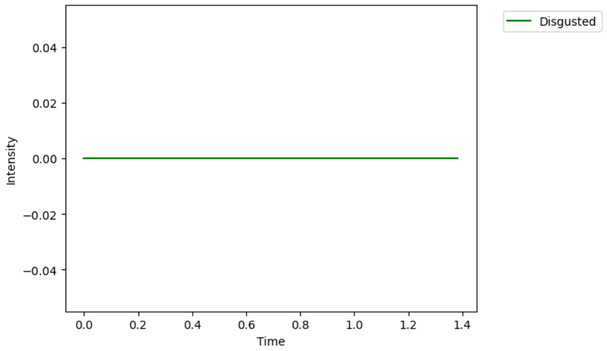
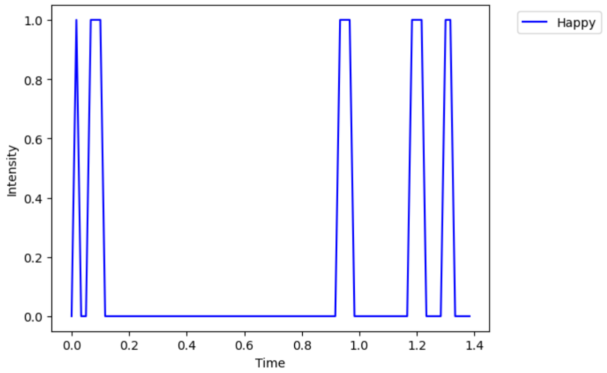


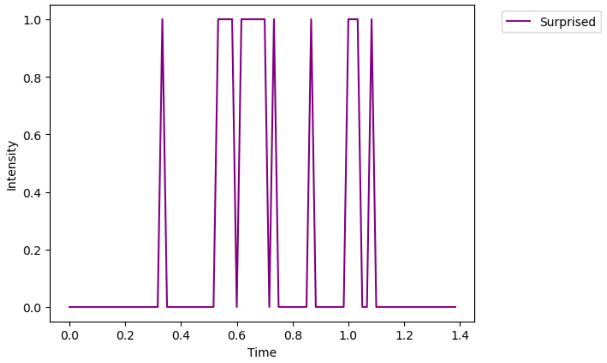
Fig. 10. Summarising reactions from live stream.











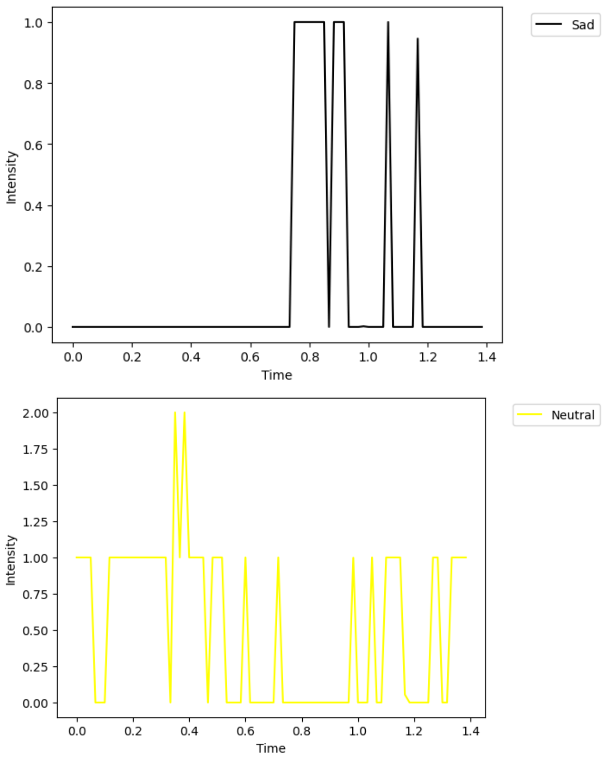


Fig. 11. Emotions Detected vs Time.

6 Conclusion

Emotion Recognition by facial expressions is one of the most challenging human-computer interaction tasks. There has been extensive research on in this topic in the past decade owing to its possibilities of numerous applications. In this project, we have trained and compared a traditional convolutional neural network model to ResNet50 transfer learning model. The results from both the models were similar denoting we achieved optimum accuracy for the dataset we used. Further research can be done to create more accurate and reliable models from a much larger dataset and compute power. We have created a system that will automatically detect the face data of individuals watching a movie or any other type of content and capture various emotions of those individuals.[25] This system will help take honest movie reviews with a good accuracy percentage. This project highlighted how current technologies and tools can used to build a primitive emotion recognition system. The system we build is capable of quantifying human facial emotions in real time and record the same. This system can be used as a tool for numerous real-world applications in various fields like entertainment, advertising, market research, psychology etc.

References

[1] Nithya Roopa.S ,“Emotion Recognition from Facial Expression using Deep Learning,” in August 2019 International Journal of Engineering and Advanced Technology,Volume-8,Issue-6S.

[2] Dilbag Singh “Human Emotion Recognition System,” in August 2012 MECS(http://www.mecs-press.org/) DOI10.5815/ijigsp.2012.08.07).

[3] Zhiwei Deng, Rajitha Navarathna, Peter Carr, Stephan Mandt, Yisong Yue, Iain Matthews, “Factorized Variational Auto encoders for Modelling Audience Reactions to Movies”, Greg Mori Simon Fraser University, Disney Research, Caltech.

[4] S. P Khandait, Dr.R.C. Thool & P.D. Khandait, “Automatic Facial Feature Extraction and Expression Recognition based on Neural Network”, (IJACSA) International Journal of Advanced Computer Science and Applications. 2, No.1, January 2011.

[5] Octavio Arriaga, Paul G. Ploger, Matias Valdenegro, “Real-time Convolutional Neural Networks for Emotion and Gender Classification”.

[6] John Gideon,Soheil Khorram,Zakaria Aldeneh,Dimitrios Dimitriadis2,Emily Mower Provost, “Progressive Neural Networks for Transfer Learning in Emotion Recognition”, University of Michigan at Ann Arbor, IBM T. J. Watson Research Centre.

[7] Prathap Nair, Andrea Cavallaro, “3-D Face Detection, Landmark Localization, and Registration Using a Point Distribution Model”, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 11, NO. 4, JUNE 2009.

[8] Jayalekshmi J, Tessy Mathew, “Facial Expression Recognition and Emotion Classification System for Sentiment Analysis”, 2017 International Conference on Networks & Advances in Computational Technologies (2017).

[9] Kamil Topal, Gultekin Ozsoyoglu, “Movie Review Analysis:Emotion Analysis of IMDb Movie Reviews”, 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[10] Anurag De , Ashim Saha, “A Comparative Study on different approaches of Real Time Human Emotion Recognition based on Facial Expression Detection”, 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA),IMS Engineering College, Ghaziabad, India.

[11] Andrew Ryan, Jeffery F. Cohn, Simon Lucey, Jason Saragih, Patrick Lucey, Fernando De la Torre, Adam Rossi, “Automated Facial Expression Recognition System”, 43rd Annual 2009 International Carnahan Conference on Security Technology,5-8 Oct. 2009.

[12] N Mehendale, “Facial Emotion Recognition using convolutional neural networks (FERC)”, SN Appl. Sci,446 (2020).

[13] C. Shetty, A. Khan, T. Singh and K. Kharatmol, "Movie Review Prediction System by Real Time Analysis of Facial Expression," 2021 6th International Conference on Communication and Electronics Systems (ICCES), (2021).

[14] Almeida, João, Luís Vilaça, Inês N. Teixeira, and Paula Viana. "Emotion Identification in Movies through Facial Expression Recognition" Applied Sciences 11, no. 15: 6827, (2021).

[15] Ijaz Ul Haq ,1 Amin Ullah ,1 Khan Muhammad ,2 Mi Young Lee ,1 and Sung Wook Baik, “Personalized Movie Summarization Using Deep CNN-Assisted Facial Expression Recognition”,

Complexity, vol. 2019, Article ID 3581419, 10 pages, (2019).

[16] Mariya A. Ali1, Dr. Sonali B. Kulkarni2, “Emotion Detection and Sentiment Analysis for Hindi Movie Reviews”, International Journal of Emerging Trends & Technology in Computer Science, Volume 10, Issue 1, (2021).

[17] Priyanka Nathawat, Dr. Vivek Chaplot, “A Review of Facial Expression Recognition”, European Journal of Molecular & Clinical Medicine, Volume 7, Issue 4, (2020).

[18] S Minaee,A Abdolrashidi, “Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network”, Sensors (SENSORS-BASEL),(2019).

[19] Singh, Chandra Bhushan and Sarkar, Babu and Yadav, Pushpendra, “Facial Expression Recognition”,SSRN, (May 25, 2021).

[20] K. Lekdioui, Y. Ruichek, R. Messoussi, Y. Chaabi and R. Touahni, "Facial expression recognition using face-regions," 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), (2017)

[21] H Nguyen, S Yeom, G Lee, H Yang, I Na and S Kim, “Facial Emotion Recognition Using an Ensemble of Multi-Level Convolutional Neural Networks”, International Journal of Pattern Recognition and Artificial Intelligence Vol. 33, No. 11, (2019).

[22] In-kyu Choi, Ha-eun Ahn and Jisang Yoo, “Facial Expression Classification Using Deep Convolutional Neural Network”, J Electr Eng Technol, (2018).

[23] Damir Filko, Prof. Goran Martinović, “Emotion Recognition System by a Neural Network Based Facial Expression Analysis”, Automatika, (2013).

[24] Milan Tripathi, “FACIAL EMOTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK”, Ictact Journal on Image and Video Processing, Volume: 12, Issue: 01, (2021).

[25] Wafa Mellouk, Wahida Handouzi, “Facial Emotion Recognition using Deep Learning: Review and Insights”, Procedia Computer Science, Volume 175, (2020).