EMOTION DETECTOR

PATTERN RECOGNITION SYSTEMS PROJECT REPORT

PRS – GROUP 13
KAH GHI LIM (A0100172A)
WEN CHENG (A0213572L)

Contents

1.0 Executive Summary

Our lives were all impacted to different extents in 2020 due to the pandemic, COVID-19. Individually, we were hit by disruptions to our daily lives due to quarantine and social distancing measures, as well as a sense of uncertainty as to when this pandemic would pass. Locally, non-essential and non-healthcare sectors such as the tourism sector was badly hit, where many were laid off. Unemployment rose and the economy contracted. Globally, many countries are still struggling to contain the spread of the pandemic, with restrictions on non-essential business travel, further contributing to the sense of uncertainty and economy contraction.

Despite this, COVID-19 accelerated digital transformation, allowing us to stay connected with our loved ones. The local government has also used digital tools for contact tracing and informing us on our potential exposure to COVID-19 infected patients. Many companies have also turned to digital solutions such as online interviews in view of the strict social distancing measures. This gives us new opportunities such as access to a global pool of talent as well as more flexible interview arrangements. Consequently, there are thus more candidates to filter and evaluate.

In this project, our group has proposed a solution, named Emotion Detector. Emotion Detector is able to automate part of the hiring process (video interview) while providing a platform for candidates to practise their video interview skills. We will also discuss limitations of the solution and suggest possible improvements for future work.

2.0 Problem Description

COVID-19 is an infectious disease caused by a newly discovered coronavirus that spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes (World Health Organization, 2020). Affecting an individual's respiratory system and being contagious, coupled with no known vaccines makes COVID-19 highly dangerous. In December 2019, the first case of COVID-19 was first reported by officials in Wuhan City, China. Close to a year after COVID-19 was first reported, there are now 39 million confirmed cases and 1.1 million deaths (WHO, 2020) globally in October 2020.

Besides affecting an individual's health negatively, COVID-19 also caused a severe economic downturn, where the Ministry of Trade and Industry of Singapore predicted setbacks to its economy, from both demand and supply in the economic survey of Singapore for the first quarter of 2020 (Saw, Lin, & Wong, 2020). Global recession and global supply chain disruptions would severely reduce demand for goods and services produced in Singapore and affect business operations respectively. By July 2020, Singapore's economy had indeed contracted 12.6% from the same period a year earlier, entering a recession in the April – June quarter (PTI, 2020).

Overall unemployment rate rose to 3.4 percent, with citizen and resident unemployment rate raising to 4.6 and 4.5 percent respectively in August (Ang, H.M., 2020), as shown in Figure 1. The Singapore government has launched many schemes to help reduce unemployment rate such as encouraging companies to retain workers through the Job Support Scheme, hiring local workers through the Jobs Growth Incentive and helping Singaporeans who are unemployed or suffered a significant income loss through the COVID-19 Support Grant (Singapore Government, 2020).

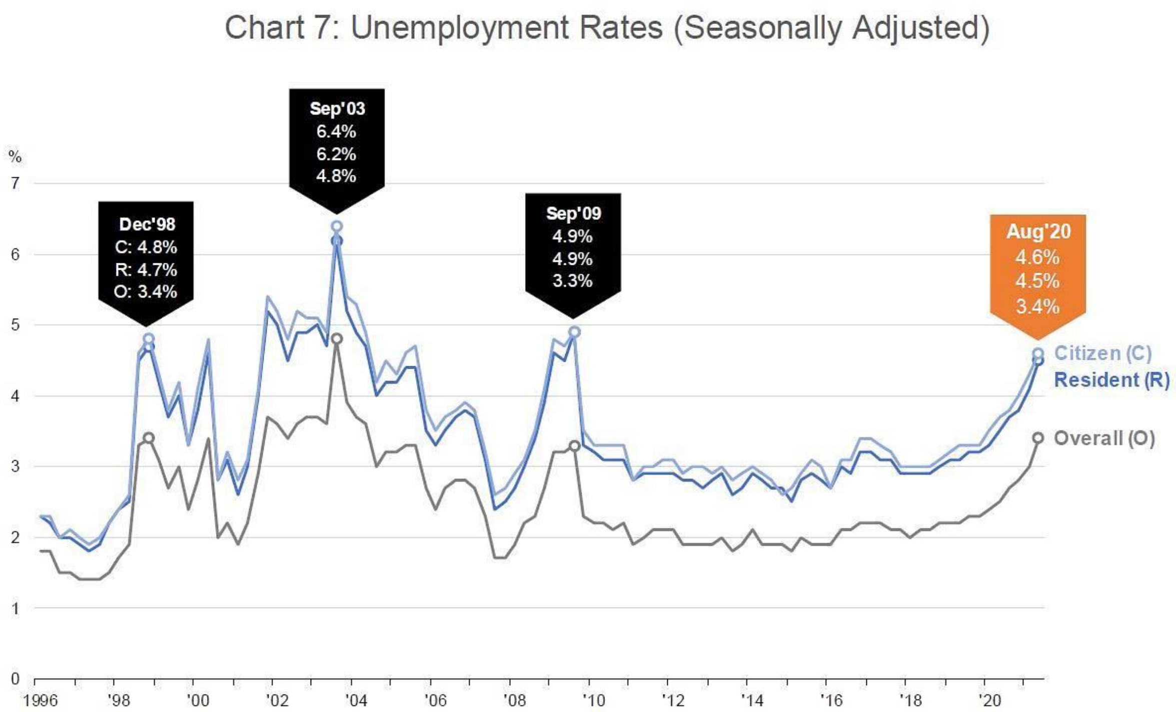


Figure 1 Unemployment rates in Singapore

With unemployment on the raise, there are now more jobseekers in the market. Coupled with the acceleration of digital transformation and COVID-19 social distancing measures, more companies have turned to digital solutions such as virtual interviews, with as many as 86% of organizations doing so (Arlington, 2020).

Virtual or video interviews not only reduces the travel to, and costs associated with getting to the office, but also gives the company an opportunity to tap into a greater pool of talent globally. This would consequently mean more work for the human resources department, where they will have to meticulously watch or attend every single video interview submitted or arranged in order to filter and shortlist potential candidates.

We thus propose our solution, Emotion Detector, where a part of the HR hiring process, namely the video interview, can be automatically analysed to give an overall impression score of the interviewee.

2.1 Project Scope and Objective

Emotion Detector is a program that recognises an individual's emotions through a screen recording, which then classifies the emotions into seven categories – happiness, neutral, sadness, anger, surprise, disgust and fear. The program would also track the duration that a particular emotion was shown and gives an overall impression score in terms of positive, neutral and negative emotion at the end of the recording.

Emotion Detector is built on a supervised learning approach, with a convolutional neural network that extracts low-level and high-level features, allowing classification of a particular emotion into seven class labels. An ensemble approach was utilised to help boost the accuracy of the model.

Emotion Detector can be used by HR for their video interviews and as a tool for interviewees to practise their interview skills at home.

3.0 Pattern Recognition System

Pattern recognition system consists of four units – a feature extractor, a pattern matcher, a reference template memory and a decision maker (Unal, 1998). The extractor selects the features required to represent the image with respect to the data label while the pattern matcher compares an input image to the reference template memory and the decision maker would then make a call on which label best describes the new input image. To build the extractor, data has to be acquired and will be elaborated in Section 3.1 while the pattern matcher, reference template memory and decision maker would be encapsulated within our system architecture in Section 4.

3.1 Data Acquisition

As part of the knowledge base, various sources of data were identified, in the form of images as shown in Table 1. Facial Expression Recognition 2013 (FER2013) was originally the only dataset we were working with. However, the highest accuracy achievable was only 69.46% and further improvements were not possible. As such, we looked to other readily available datasets to increase and enhance the database for the neural network to learn from. We decided on the Extended Cohn-Kanade (CK+) (Lucey, et al) and a facial expression dataset obtained from muxspace GitHub repository.

In total, the combined dataset contains 50,494 samples of 48x48 pixel grayscale images, after removing samples with missing labels and labelled contempt to avoid an imbalanced dataset.

Table 1: Sources of data

No.	Data source	Details
1	Facial Expression Recognition 2013 (FER2013)	35,887 examples 48x48 pixel grayscale images 7 classes: happiness, neutral, sadness, anger, surprise, disgust, fear
2	Extended Cohn-Kanade Dataset (CK+)	981 examples 48x48 pixel 7 classes: happy, sadness, anger, surprise, disgust, fear, contempt

3	Facial expressions dataset from muxspace github (https://github.com/muxspace/facial_expressions)	13,717 images 350x350 pixel 7 classes: happiness, neutral, sadness, anger, surprise, disgust, contempt

3.2 Data Model

The images were all converted to grayscale prior to training as we are more interested in the facial expression and not other features like skin colour. Images were also augmented during training to account for movement during the video interview, such as width and height shift, rotation and zoom range. Horizontal flip was also set true to account for flipped images while vertical flip was set false as we do not expect an upside-down image during a video interview.

Examples of data augmentation are shown in Figure 2.

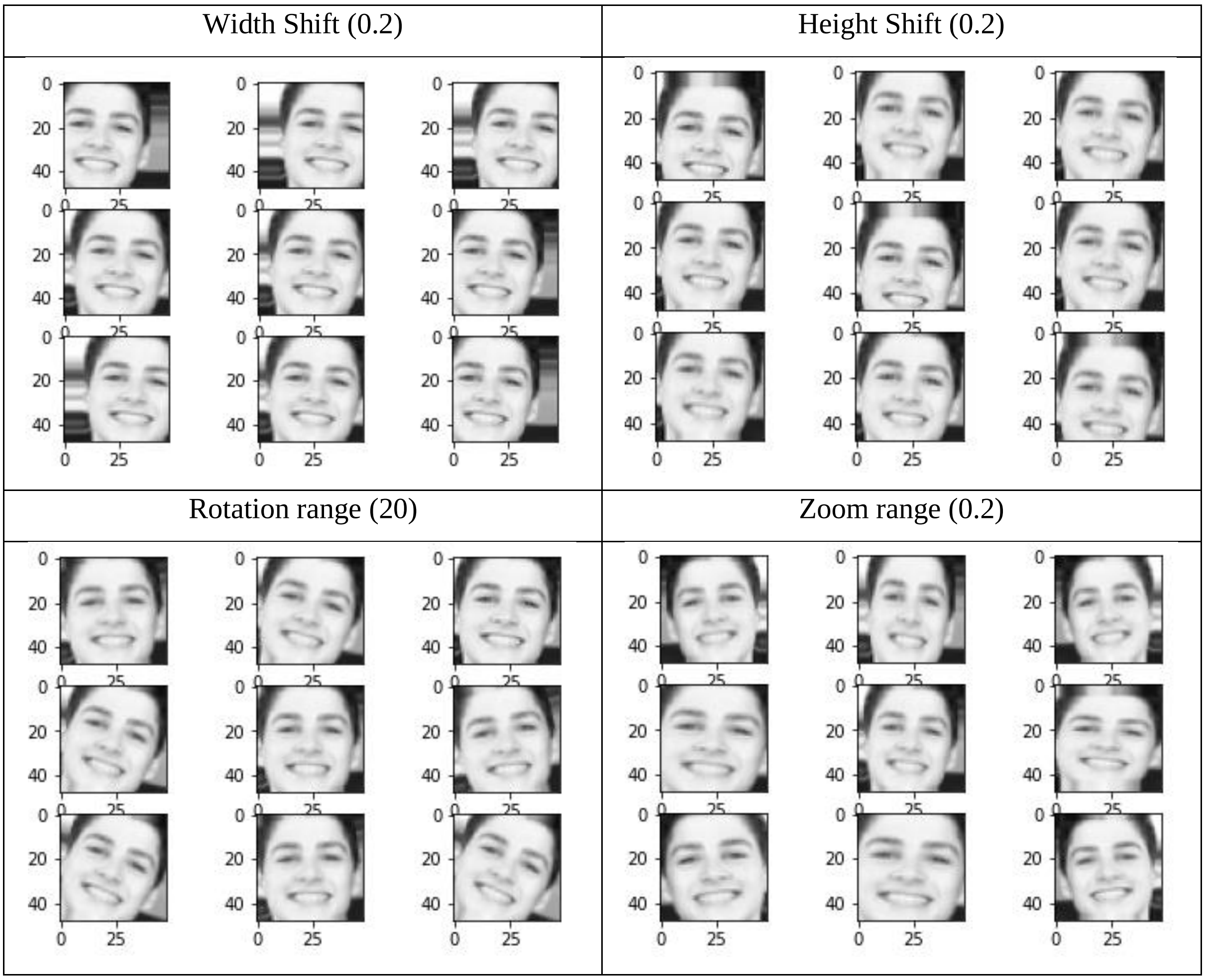


Figure 2 Examples of Data Augmentation

3.3 Neural Network Structure

Various network architectures were explored, with a summary of accuracies achieved as shown in Table 2. Techniques such as He initializer, batch normalization, kernel initializers and dropout layers were utilized to prevent overfitting of the model. More details on the F1- score and loss values for each training run can be found in Appendix.

Table 2: Accuracies achieved with different models and different datasets

No.	Dataset used	Model	Details	Accuracy (%)
1	FER2013	ConvNet	100 epochs	67.30
2	FER2013	ConvNet	150 epochs He initializer	69.14
3	FER2013	ConvNet	200 epochs He initializer	69.46
4	FER2013	ResNet50	50 epochs Learning rate scheduler	55.38
5	FER2013	VGG13	50 epochs Learning rate scheduler	66.65
6	FER2013	VGG19	60 epochs He initializer Kernel regularizer Dropout Learning rate scheduler	69.35
8	FER2013, CK+ and muxspace	ConvNet	200 epochs He initializer Kernel regularizer	73.27
9	FER2013, CK+ and muxspace	3 ConvNet, voting ensemble	200 epochs He initializer Kernel regularizer	72.86
10	FER2013, CK+ and muxspace	VGG19	60 epochs He initializer Kernel regularizer Dropout Learning rate scheduler	73.23
11	FER2013, CK+ and muxspace	3 VGG19, voting ensemble	60 epochs He initializer Kernel regularizer Dropout Learning rate scheduler	73.37

Using only the FER2013 dataset, the two models that had the highest accuracy were ConvNet (69.46%) and VGG19 (69.35%). This result is comparable to the second place in the Kaggle competition ‘Challenges in Representation Learning: Facial Expression Recognition Challenge’ (Kaggle, 2013).

To further increase the accuracy of our model, the combined dataset consisting of FER2013, CK+ and muxspace was used to retrain the model. This further increased the accuracy to 73.27% and 73.23% for ConvNet and VGG19 respectively.

Finally, to boost the accuracy further, a voting ensemble approach was utilised. 3 separate models were trained with different He initializers and each model will predict the outcome separately. Finally, majority vote is used as the final prediction, yielding 72.86% and 73.37% accuracy for ConvNet and VGG19 respectively. The final model is then taken as 3 VGG19 models, with voting ensemble since this gave the highest accuracy. Figure 3 shows the architecture for a single VGG19 model.

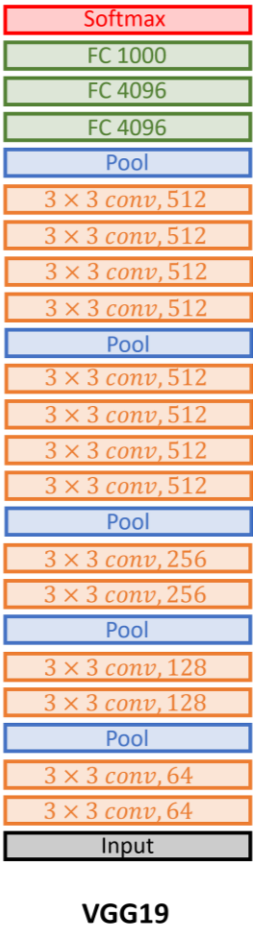


Figure 3 VGG19 architecture

From the training of the neural network, it can be seen that the quality and quantity of the dataset influences the accuracy of the trained model. Before the introduction of CK+ and muxspace dataset into the model, the accuracy of the model would only achieve a maximum of ~69% and was not able to breach the 70% mark. This could be due to poorly labelled data within the dataset, thus limiting the achievable accuracy. Although the accuracy of the model increased with the introduction of the CK+ and muxspace dataset, indicating that the other two datasets may be better labelled, we did not remove the FER2013 dataset as more than half the dataset would be removed.

4.0 Solution

4.1 Solution Overview

To help the user collect and analyse the emotion data of the interviewee, our application will grab part of the screen as input where user may place the interview window. Once user starts the session, the application will read the interview screen in which the interviewee's face must be presented. If face is not detected, there will be a warning message on the UI to notify the user.

The face data will then be passed to the ensemble CNN model to make the emotion prediction and a data logger will log the total time duration for each emotion. After user end the session, a graph with summarized emotion data will be presented to the user and he will be able to save it for further analysis.

4.2 System Architecture

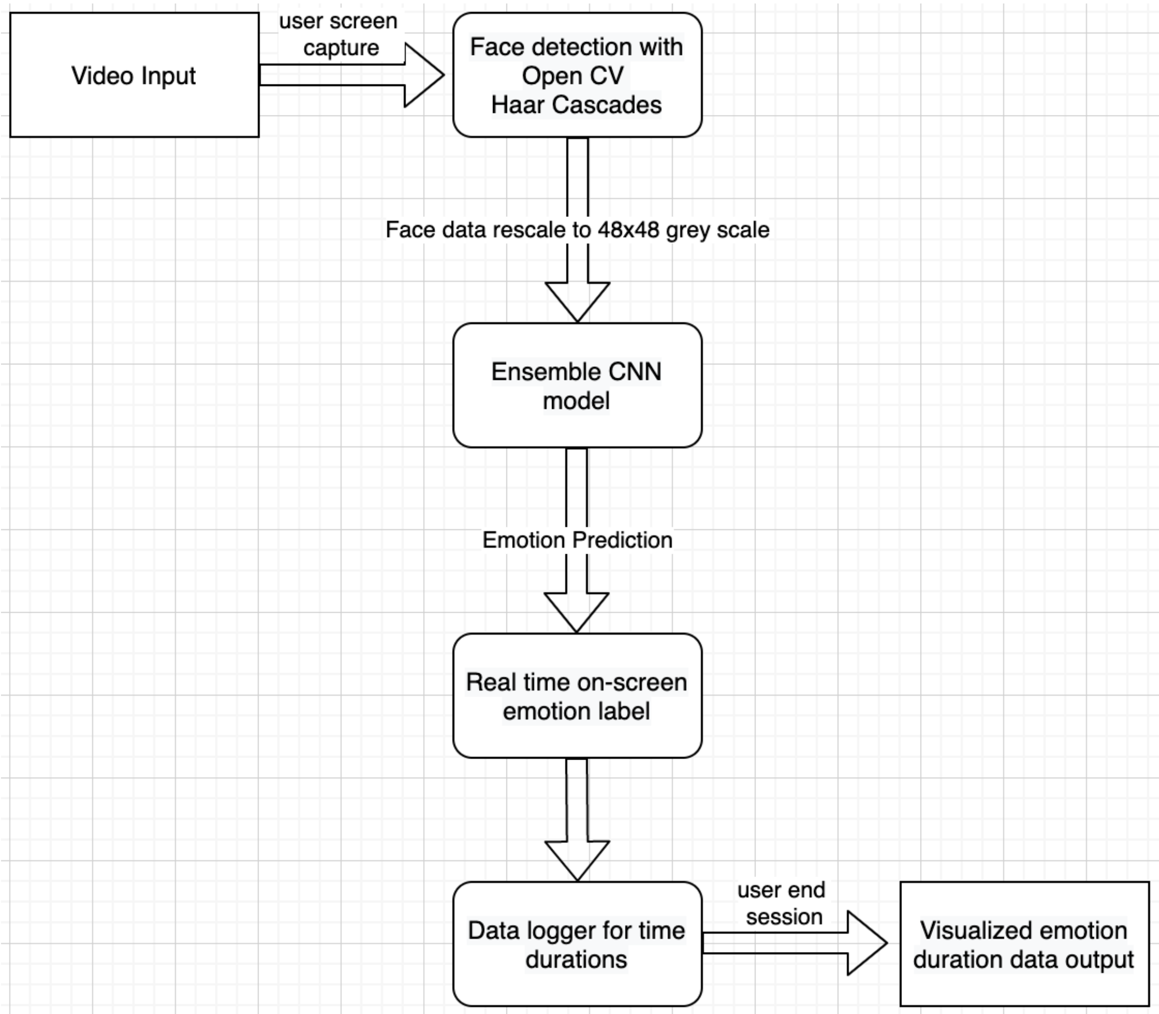


Figure 4 System Architecture Diagram

4.3 System Flow

The system architecture diagram shown in Figure 4 demonstrates the overall flow of the application. When user starts the interview, the window will be captured as video input and feed to the application backend for processing. For the face detection mechanism, the application used Open CV Haar feature based cascade classifiers to identify the face as well as the bounding box. The face data is then rescaled to 48x48 grey scale image and feed to the ensemble CNN model to make the emotion prediction. The final prediction result will be the class with the highest probability and the application will further categorize the predicted emotion into Positive/Neutral/Negative to give the user a summarized result.

The application will also display the predicted emotion category in real time on the screen so the user can monitor the interviewee's emotion throughout the session. Once the user clicks the 'End' button, a bar chart for the duration for each emotion category will be displayed. The bar for each category will also be colour coded with a label of the duration in seconds to give the user the best visualization of the final result. An example is shown in Figure 5, with green, grey and red bars for Positive, Neutral and Negative emotion respectively. User may save the graph for further analysis.

Figure 5 Summarized bar chart once application is ended

4.4 Assumptions

Some of the key underlying assumptions are:

1. There should be only one interviewee at a time in the window
2. The interview window must be placed at a certain place of the screen

4.5 System Features

The key features of our application are

- Standalone executable application (Windows)
- Intuitive UI
- Real-time emotion detection and labelling
- Data logger and visualized output

To run our application on Windows platform is very simple, it can be launched by the executable file without any development environment required. The design of the UI is also intuitive, user only need to click the toggle button to start and end the session.

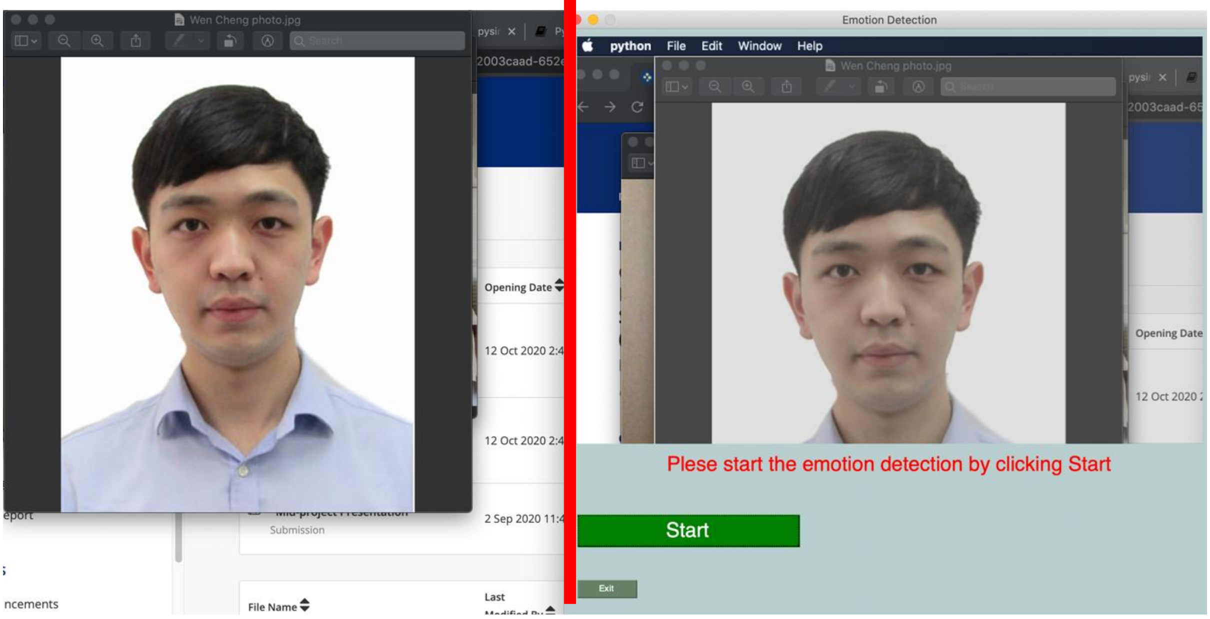


Figure 6 Application before start

By default, the application will run on the right half of the screen and it will capture everything on the top left of screen as illustrated in Figure 6 and 7. User will place the interview window on the top left of the screen for the program to capture. Once ready, user can click Start button and the program will start to predict and record the emotion data and the real time emotion will be displayed in the UI.

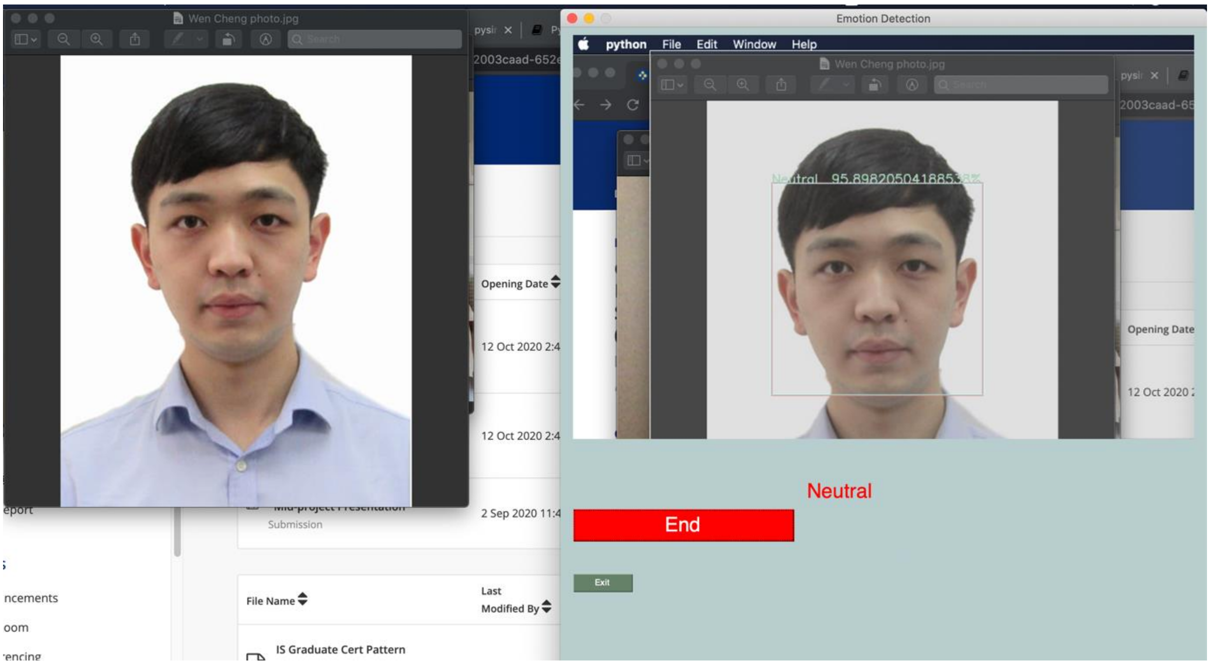


Figure 7 Application during recording process

During the session, if the program fails to identify any face data, it will prompt a warning to notify the user as shown in Figure 8.

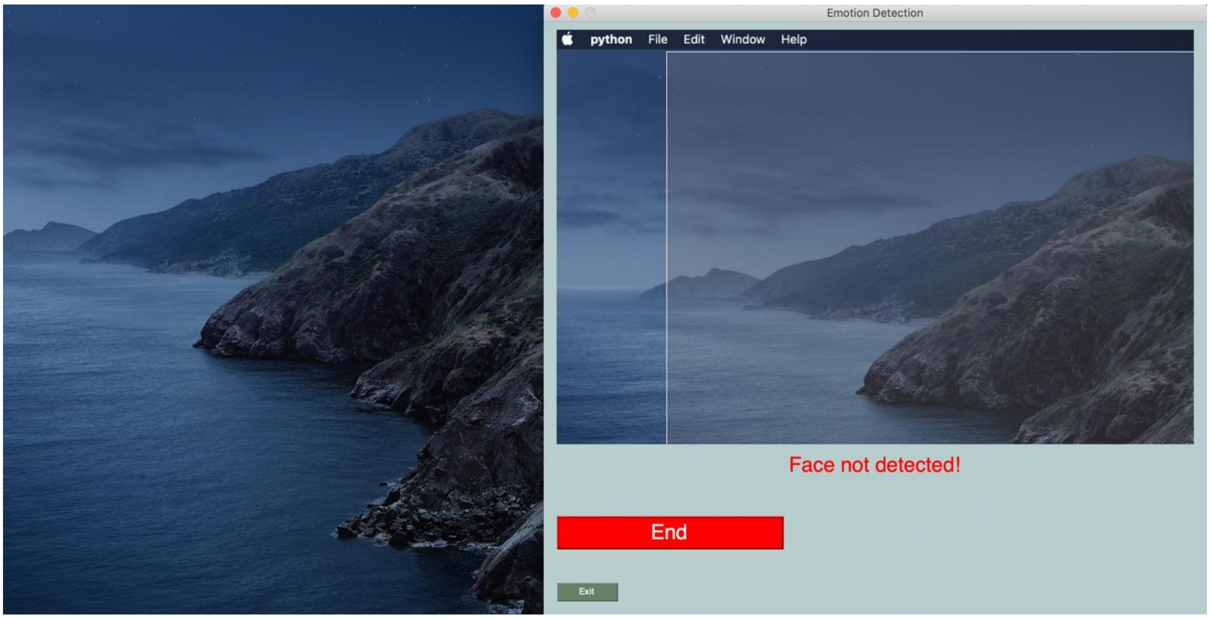


Figure 8 Error message if face is not detected

When user clicks the End button, the program will generate the summarized bar chart and prompt to the user as shown in Figure 9. User may save the graph or they can start another session for a new interviewee without restarting the program.

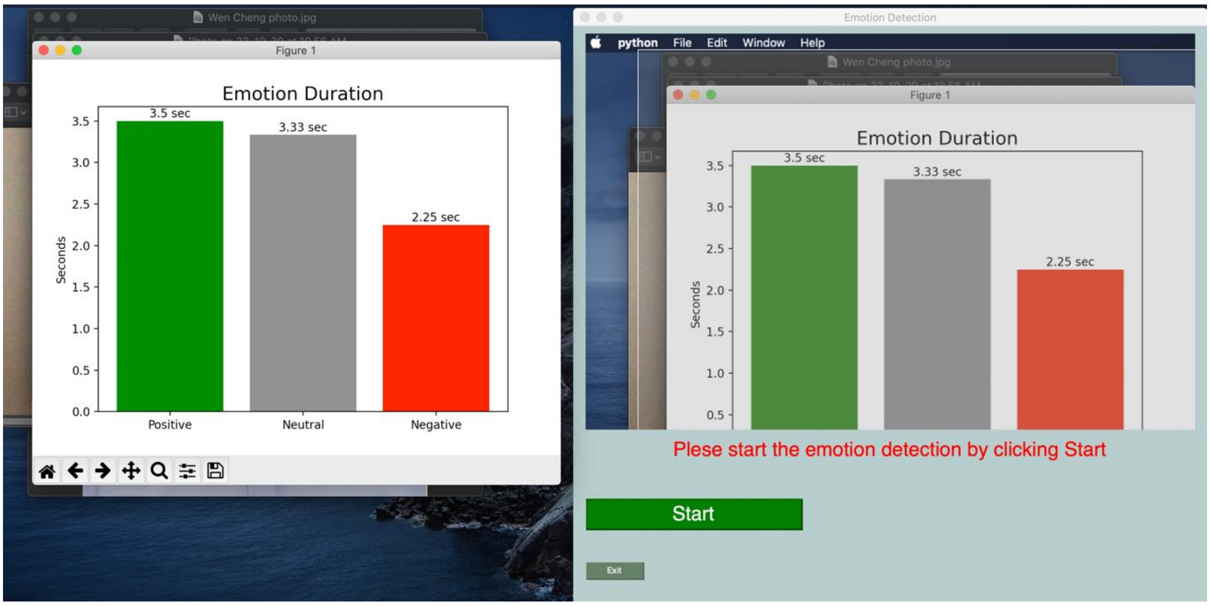


Figure 9 Application upon ending the recording

4.6 Limitations

The biggest limitation of our application is that it only allows one person be present in the interview window. If there are multiple interviewees, the program will not be able to tell the emotion for different person. Another limitation is that in order to run the application on non-Windows platform (Linux, MacOS), it will require a development environment with Python and other necessary libraries installed which will be difficult for users without any technical background to set up.

5.0 Conclusions & References

The outbreak of COVID-19 around the world certainly creates a new norm in the global society where more work and social interactions will be conducted online. This gives a lot of opportunities for our application to be used, where the use case can be extended to group meetings, conducting online courses and many other aspects. There is definitely a lot of potential market value in our product.

There is also room for future improvement as well, the application can be integrated with any online communication apps like Zoom, Skype and Teams as a useful feature. Furthermore, we can also analyse the audio signal of the interviewee to determine how engaged or excited the interviewee sounds, thereby increasing the accuracy for the prediction of the emotion displayed.

6.0 Bibliography

Ang, H. M. (2020, Oct 07). Singapore's overall unemployment rate in August climbs past global financial crisis' high. *Channel News Asia*. Retrieved from https://www.channelnewsasia.com/news/singapore/unemployment-rate-singapore-august-covid-19-13218836

Arlington, V. (2020,Apr 30). Gartner HR Survey Shows 86% of Organizations Are Conducting Virtual Interviews to Hire Candidates During Coronavirus Pandemic. *Gartner*. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2020-04-30-gartner-hr-survey-shows-86--of-organizations-are-cond

Kaggle. (2013). *Challenges in Representation Learning: Facial Expression Recognition Challenge.* Retrieved from https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/leaderboard

Lucey, et al. *The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression.* Retrieved from http://www.jeffcohn.net/wp-content/uploads/2020/02/CVPR2010_CK2.pdf.pdf

PTI. (2020, Jul 14) COVID-19 impact: Singapore's economy contracts 41.2% on quarterly basis in April-June. *Economic Times.* Retrieved from https://economictimes.indiatimes.com/news/international/business/covid-19-impact-singapores-economy-contracts-41-2-on-quarterly-basis-in-april-june/articleshow/76960695.cms

Saw, C., Lin, J. & Wong, Y.J. (2020). Feature Article: Impact of the COVID-19 Pandemic on the Singapore Economy. *Ministry of Trade and Industry*. Retrieved from https://www.mti.gov.sg/-/media/MTI/Resources/Economic-Survey-of-Singapore/2020/Economic-Survey-of-Singapore-First-Quarter-2020/FA_1Q20.pdf

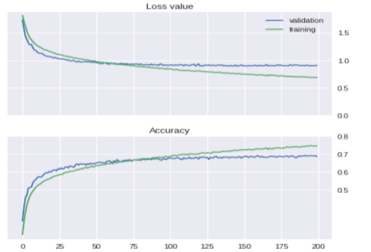
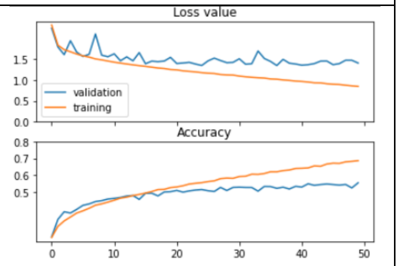
Singapore Government. (2020). *More support for workers and jobs through the Jobs Support Scheme and COVID-19 Support Grant*. Retrieved from https://www.gov.sg/article/more-support-for-workers-and-jobs-through-the-jobs-support-scheme-and-covid-19-support-grant#:~:text=The%20Jobs%20Support%20Scheme%20(JSS,paid%20up%20to%20March%202021.

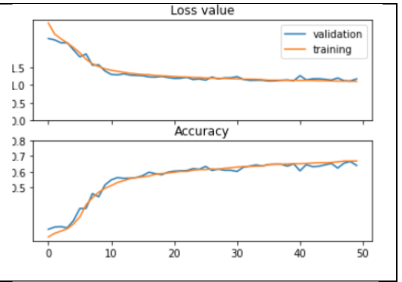
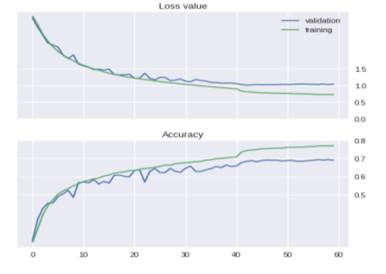
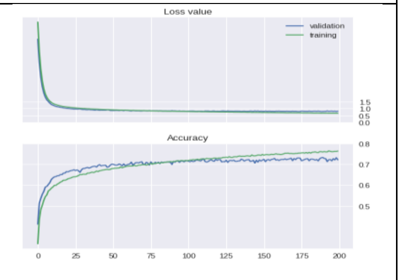
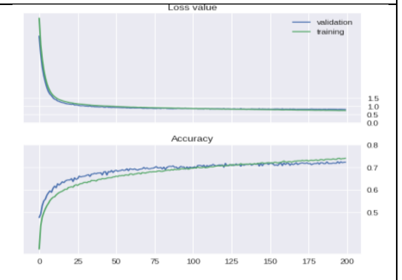
Unal, F.A. (1998). Citation: Temporal Pattern Matching Using an Artificial Neural Network. ScienceDirect. Retrieved from https://www.sciencedirect.com/topics/engineering/pattern-recognition-systems

World Health Organization. (2020). *Coronavirus.* Retrieved from https://www.who.int/health-topics/coronavirus#tab=tab_1

World Health Organization. (2020). *WHO Coronavirus Disease (COVID-19) Dashboard.* Retrieved from https://covid19.who.int/

7.0 Appendix

No.	Dataset used	Model	Details	F1-score	Loss curves
1	FER2013	ConvNet	100 epochs	Best accuracy (on testing dataset): 67.30% precision recall f1-score support angry 0.5934 0.6033 0.5983 958 disgusted 0.8261 0.5135 0.6333 111 fearful 0.5484 0.3926 0.4576 1024 happy 0.8528 0.8850 0.8686 1774 neutral 0.5930 0.7113 0.6468 1233 sad 0.5580 0.5477 0.5528 1247 surprised 0.7739 0.7990 0.7863 831 accuracy 0.6730 7178 macro avg 0.6779 0.6361 0.6491 7178 weighted avg 0.6694 0.6730 0.6678 7178	
2	FER2013	ConvNet	150 epochs He initializer	Best accuracy (on testing dataset): 69.14% precision recall f1-score support angry 0.6646 0.5564 0.6057 958 disgusted 0.7848 0.5586 0.6526 111 fearful 0.5764 0.4570 0.5098 1024 happy 0.8812 0.8861 0.8836 1774 neutral 0.5845 0.7486 0.6565 1233 sad 0.5831 0.5597 0.5712 1247 surprised 0.7643 0.8508 0.8052 831 accuracy 0.6914 7178 macro avg 0.6913 0.6596 0.6692 7178 weighted avg 0.6910 0.6914 0.6873 7178	
3	FER2013	ConvNet	200 epochs He initializer	Best accuracy (on testing dataset): 69.46% precision recall f1-score support angry 0.6175 0.6200 0.6188 958 disgusted 0.8289 0.5676 0.6738 111 fearful 0.5993 0.4717 0.5279 1024 happy 0.8756 0.8890 0.8822 1774 neutral 0.6040 0.7088 0.6522 1233 sad 0.5695 0.5686 0.5690 1247 surprised 0.8157 0.8255 0.8206 831 accuracy 0.6946 7178 macro avg 0.7015 0.6645 0.6778 7178 weighted avg 0.6942 0.6946 0.6922 7178	
4	FER2013	ResNet50	50 epochs Learning rate scheduler	Best accuracy (on testing dataset): 55.38% precision recall f1-score support angry 0.4714 0.4301 0.4498 958 disgusted 0.5833 0.5045 0.5411 111 fearful 0.4272 0.4238 0.4255 1024 happy 0.7321 0.7593 0.7454 1774 neutral 0.4390 0.5839 0.5012 1233 sad 0.4610 0.3881 0.4214 1247 surprised 0.7885 0.6282 0.6993 831 accuracy 0.5538 7178 macro avg 0.5575 0.5311 0.5405 7178 weighted avg 0.5606 0.5538 0.5536 7178	

5	FER2013	VGG13	50 epochs Learning rate scheduler	Best accuracy (on testing dataset): 66.65% precision recall f1-score support angry 0.6122 0.5981 0.6051 958 disgusted 0.7391 0.4595 0.5667 111 fearful 0.6097 0.2959 0.3984 1024 happy 0.8476 0.8844 0.8657 1774 neutral 0.5686 0.7226 0.6364 1233 sad 0.5312 0.5878 0.5581 1247 surprised 0.7563 0.7990 0.7771 831 accuracy 0.6665 7178 macro avg 0.6664 0.6211 0.6296 7178 weighted avg 0.6671 0.6665 0.6565 7178	
6	FER2013	VGG19	60 epochs He initializer Kernel regularizer Dropout Learning rate scheduler	Best accuracy (on testing dataset): 69.35% precision recall f1-score support Angry 0.6389 0.6075 0.6228 958 Disgust 0.7590 0.5676 0.6495 111 Fear 0.5829 0.4668 0.5184 1024 Happy 0.8860 0.8766 0.8813 1774 Sad 0.5906 0.5750 0.5827 1247 Surprise 0.7834 0.8267 0.8044 831 Neutral 0.5903 0.7267 0.6514 1233 accuracy 0.6935 7178 macro avg 0.6902 0.6638 0.6729 7178 weighted avg 0.6938 0.6935 0.6912 7178	
8	FER2013, CK+ and muxspace	ConvNet	200 epochs He initializer Kernel regularizer	Best accuracy (on testing dataset): 73.27% precision recall f1-score support Angry 0.6366 0.5794 0.6067 1070 Disgust 0.7987 0.6287 0.7036 202 Fear 0.5658 0.4741 0.5159 1061 Happy 0.8772 0.9012 0.8890 2916 Sad 0.5772 0.5318 0.5536 1307 Surprise 0.7308 0.7993 0.7635 917 Neutral 0.7286 0.7974 0.7615 2626 accuracy 0.7327 10099 macro avg 0.7021 0.6731 0.6848 10099 weighted avg 0.7267 0.7327 0.7282 10099	
9	FER2013, CK+ and muxspace	3 ConvNet, voting ensemble	200 epochs He initializer Kernel regularizer	Best accuracy (on testing dataset): 72.86% precision recall f1-score support Angry 0.5588 0.6393 0.5963 1070 Disgust 0.8417 0.5792 0.6862 202 Fear 0.5868 0.4458 0.5067 1061 Happy 0.8817 0.9026 0.8921 2916 Sad 0.5889 0.5042 0.5433 1307 Surprise 0.7492 0.7917 0.7699 917 Neutral 0.7235 0.7871 0.7540 2626 accuracy 0.7286 10099 macro avg 0.7044 0.6643 0.6783 10099 weighted avg 0.7247 0.7286 0.7240 10099	

10	FER2013, CK+ and muxspace	VGG19	60 epochs He initializer Kernel regularizer Dropout Learning rate scheduler	Best accuracy (on testing dataset): 73.23% precision recall f1-score support Angry 0.5777 0.6150 0.5957 1070 Disgust 0.7746 0.6634 0.7147 202 Fear 0.5818 0.4995 0.5375 1061 Happy 0.8952 0.8995 0.8974 2916 Sad 0.5587 0.5830 0.5706 1307 Surprise 0.7734 0.7666 0.7700 917 Neutral 0.7430 0.7563 0.7496 2626 accuracy 0.7323 10099 macro avg 0.7006 0.6833 0.6908 10099 weighted avg 0.7320 0.7323 0.7317 10099	
11	FER2013, CK+ and muxspace	3 VGG19, voting ensemble	60 epochs He initializer Kernel regularizer Dropout Learning rate scheduler	Best accuracy (on testing dataset): 73.37% precision recall f1-score support Angry 0.5447 0.6486 0.5922 1070 Disgust 0.8302 0.6535 0.7313 202 Fear 0.5906 0.4637 0.5195 1061 Happy 0.8942 0.9074 0.9008 2916 Sad 0.5693 0.5501 0.5595 1307 Surprise 0.7975 0.7688 0.7829 917 Neutral 0.7415 0.7700 0.7555 2626 accuracy 0.7337 10099 macro avg 0.7097 0.6803 0.6917 10099 weighted avg 0.7335 0.7337 0.7320 10099