MicroX Emotion Recognition

Iman Heizer Pua
Information Systems Technology and Design
Singapore University of Technology and Design (SUTD)
Singapore
heizerpua_iman@mymail.sutd.edu.sg

Jazreel Kwek
Information Systems Technology and Design
Singapore University of Technology and Design (SUTD)
Singapore
jazreel_kwek@mymail.sutd.edu.sg

Melissa
Information Systems Technology and Design
Singapore University of Technology and Design (SUTD)
Singapore
melissa@mymail.sutd.edu.sg

Ryan Gen Zi Qiang
Information Systems Technology and Design
Singapore University of Technology and Design (SUTD)
Singapore
ryan_gen@mymail.sutd.edu.sg

***Abstract*—Deep Neural Networks (DNN) has much potential in a variety of use cases. Emotion recognition is a key feature in the field of robotics, sales in determining customer satisfaction and even self-driving cars to look out for driver state through driver monitoring. The challenge of emotion detection is now greater during this Covid-19 period with individuals using face masks. We demonstrate the use of Convolutional Long Short Term Memory models against video inputs for this classification problem and how using multiple blocks containing different micro expression features can increase the accuracy of the model in diverse scenarios (Person with mask on).**

I. Introduction

Convolutional neural networks have made remarkable breakthroughs in the field of computer vision, allowing for the basic feature extractions to be optimized, so that patterns between different pixels within an image can be taken into account as the image is fed through the neural network.

While most studies focus on the model and its parameters, our study aims to add an additional depth to the field of emotion recognition, by learning from the best applications of neural networks and processing, us humans. As we try so hard to innovate and derive new models, our study places emphasis instead on the approach, and how we intend to break down the problem. Studies of the human behaviour, body language and micro expressions specialists are those who we can learn from. This forms the motivation for our micro expression (MicroX) model. While the average human understands emotion based on past exposure or experience, subconsciously being able to tell how another person might be feeling through his/her facial expressions, experts on the other hand are able to quantify what they see before them and make accurate judgements. This source code can be found at this Github link https://github.com/HeizerSpider/MicroX_Emotion_Recognition.git

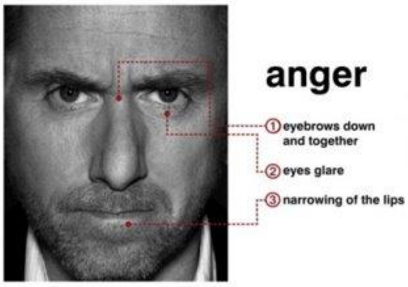


Figure 1: Micro expression breakdown

II. Dataset Collection

A. BAUM-1 Datasets

Collection of audio-visual facial clips of acted and spontaneous expressions (273, 1184 clips respectively). These clips were recorded from 31 individuals, in front of green screens and with studio lighting.

B. BAUM-2 Datasets

Collection of audio-visual facial clips extracted semi-automatically from movies and TV series (1047 clips). These clips were from diverse conditions.

We had access to the IEMOCAP dataset, however we had troubles downloading the huge files due to memory constraints (12 and 17GB), so we were unable to use it. We also had access to the Enterface dataset, but because the file structure and labeling was completely different from BAUM datasets, we decided not to use it either. Other datasets that we requested permission for access were RAVDESS, Belfast Database, MMI Database, but were unavailable. The BAUM datasets come in similar formats. They are folders of different video sets, and an accompanying csv containing the labels of all videos in the dataset.

Augmentation of the data to increase the size of our dataset was also considered but given that there was enough data given the amount of time left to train the model, hence we did not carry on with the process. This could be a future implementation to further improve the model.

III. Pre-processing

A. Data Generation

From the raw dataset of videos, the videos are extracted into frames using openCV with a frame per second(fps) rate of 30fps. These frames are then fed into the OpenFace Model[1], we use the facial landmark detection submodule to get the pixel coordinates of key facial parts, what we call MicroX components (Short for micro expressions, which includes the 1) Left eyebrow, 2) Right eyebrow, 3) Left eye, 4) Right eye, 5) Nose and 6) Mouth). These coordinates are needed to be able to split the respective frames into different parts representative of each facial feature.

In the current iteration of the trained model however, due to the low accuracy of the coordinates given for the left and right eyebrow, we have chosen to omit the processed MicroX data for these parts. The csv generated from OpenFace however was not entirely suitable for our use due to the format of the data and hence required us to extract only what was needed (eg. pixel coordinates) and include data such as frame name into the csv for us to use.

With the pixel coordinates, we are then able to crop out and resize the MicroX components using the openCV library once more so that they will have fixed dimensions of 30 by 30 pixels to be fed into our model. Each of the frames will thus have 4 MicroX components each and so at any one point in time, 10 frames of 4 MicroX (40 images) together with the full-face frames (resized to 30 by 30 as well) will be prepared to be fed into the model.

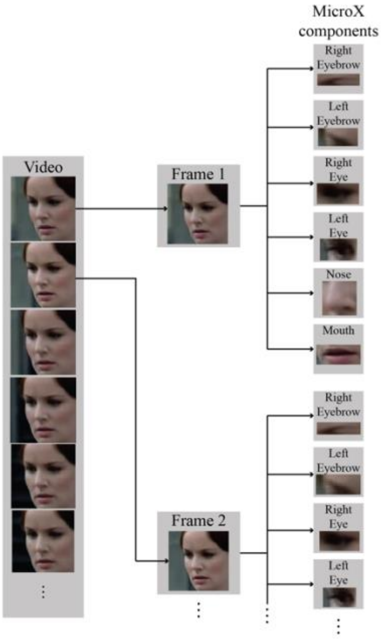


Figure 2: Pre-processing of full-face image

B. *Pipelining*

Further processing was needed to increase the efficiency and ensure a smooth pipeline flow across the datasets we were using. Data cleaning was thus necessary to ensure that all the data the model received were of the same standard, such as no frames, and that the csv generated was not corrupted. If this was not done, there would be many inconsistencies in the model training which would have slowed down the overall progress as well as accuracy of the model.

Each raw unprocessed video also generated different number of frames ranging from 20 to 400 frames depending on the video duration. In order to maximize the use of all the frames rather than limiting it across all videos, segmenting of the video into groups of 10 frames was necessary. This potentially increased our dataset by a great amount which would be crucial in getting a better model outcome.

The csv containing labelled data of the different videos and frames was also adjusted to fit our usage. Although the different datasets (BAUM-1 and BAUM-2) were used to train models for emotion recognition, they had different class labels such as BAUM-1 taking into consideration a few more emotions as compared to BAUM-2.

Emotions for BAUM-1 dataset:

{Anger, Boredom, Bothered, Concentrating, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, Surprise, Thinking, Unsure, Interest}

Emotions for BAUM-2 dataset:

{Neutral, Angry, Contempt, Disgust, Fear, Happy, Sadness, Surprise}

As such, we aggregated this data to limit it to 7 distinct classes which allow us to keep the output classes across both datasets the same. An example of such a change would be combining the class "Contempt" and "Disgusting" since both were very similar. As such, the final dataset was a combination of both BAUM-1 and BAUM-2 with the emotion classes following that of BAUM-2. Figure 3 shows the final distribution of data across the different classes after concatenation of the 2 datasets.

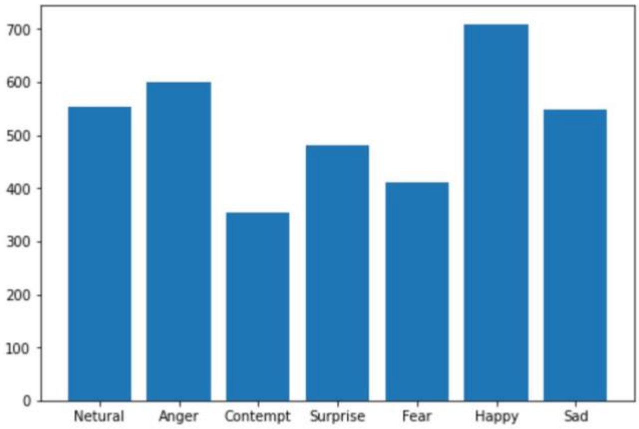


Figure 3: Data Distribution

IV. Problem and Algorithm/Model Evaluation Methodology

A. *Problem*

With the proliferation of video-sharing platforms such as social media networks and video conferencing tools, video-based, sentiment analysis is increasingly becoming more popular due to its numerous applications. Sentiment analysis from videos is currently widely used for a variety of use cases such as customer satisfaction, brand perception analysis, future recommendations, driver monitoring in Advanced Driver Assistance Systems, among other things. What might be lacking in recent development in video-based sentiment analysis is the lack of focus in micro expressions. By taking a micro expression-based approach, our model can leverage emotion recognition from each micro feature to determine an individual's sentiment even with less data such as not having full access to the person's face.

For our purpose, we will be using the ConvLSTM cell which is a variation of LSTM cell that performs convolution within the LSTM cell. Compared to LSTM cell, it replaces the matrix multiplication with the convolution operation. This will be desirable for our video sentiment analysis classification task as it captures the spatial features from the image.

B. *Model*

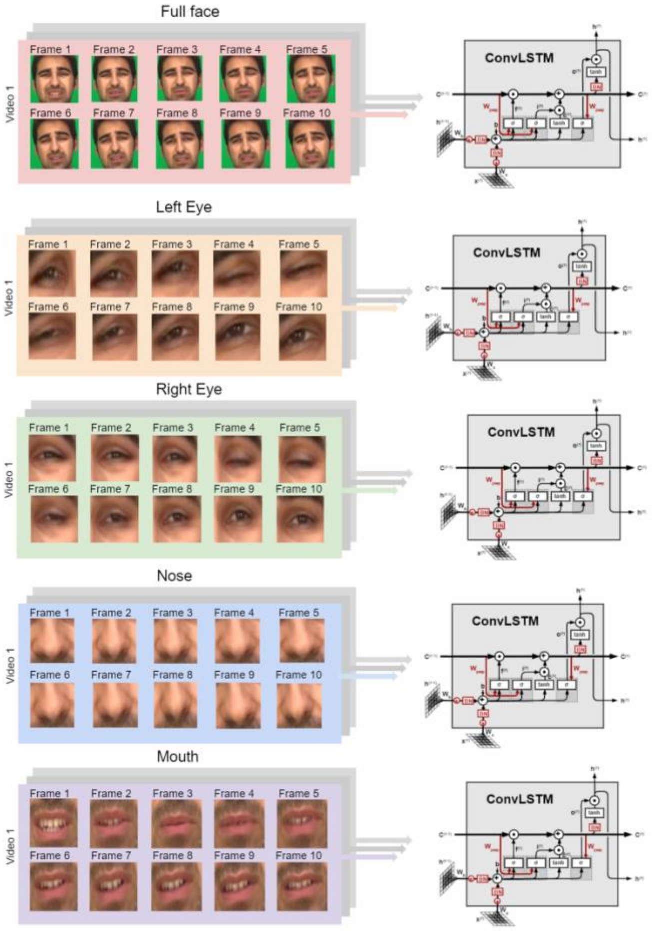


Figure 4: Multiple Convolutional LSTM blocks

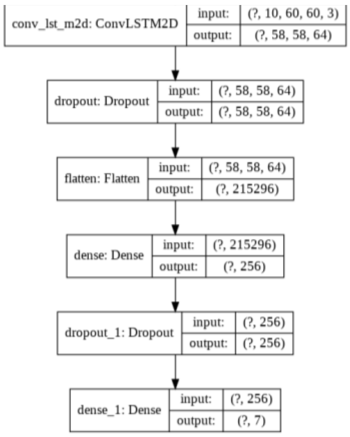
Our model takes into consideration temporal images derived from frames in sequence, with each input spanning 10 frames from a single video. Frames are extracted from videos using OpenCV with a frame rate of 30fps. For each of these frames, we further break them down into its individual micro expression components. Mainly, the right eye, left eye, nose and mouth. We also retain the original full-face frame. We then feed each of these components into a single ConvLSTM block. This is done so that each MicroX component will output a value in relation to time that is dependent only on the component itself (eg. The changes in the right eye relative to time will be given an output value that is independent of the other MicroX components).

C. *ConvLSTM*

Before feeding the frames into the cell sequentially, we standardized the height and width of these individual images and then extracted the RGB values of each picture. For each of these blocks, we set some default parameters:

- Filters = 64
- Kernel_size = (3,3)
- return_sequence = False
- Input_shape = (10,60,60,3)

Parameter return_sequence is set to be False as we are not interested in the output at every time step but rather the predicted emotion after the final timestep. After the ConvLSTM layer, we apply a dropout layer of 0.2 for regularization to prevent overfitting. The output is then flattened to be fed into a dense layer with 256 neurons. Relu activation function is used to introduce non-linearity into the neural network since images are naturally non-linear. Another dropout layer is then added before feeding it into the final dense layer with 7 neurons to get the predicted output. SoftMax activation function is used for this final dense layer as we are dealing with a multi-class classification problem. Finally, we get the output (predicted emotion) for each MicroX. We repeat this process for all the MicroX we wish to consider as well as the full face.



conv_lst_m2d: ConvLSTM2D	input:	(?, 10, 60, 60, 3)
	output:	(?, 58, 58, 64)
dropout: Dropout	input:	(?, 58, 58, 64)
	output:	(?, 58, 58, 64)
flatten: Flatten	input:	(?, 58, 58, 64)
	output:	(?, 215296)
dense: Dense	input:	(?, 215296)
	output:	(?, 256)
dropout_1: Dropout	input:	(?, 256)
	output:	(?, 256)
dense_1: Dense	input:	(?, 256)
	output:	(?, 7)

Figure 5: Model Summary

D. *Functional API*

The output from each of these ConvLSTM blocks will then be used as input to the final layer. Since we have 6 separate cell blocks, we now have 6 different layers. Through a functional API, we concatenate these inputs and return a single tensor that concatenates all the inputs. We feed this output into a final dense layer to find the best weighted sum to predict the final emotion. We now have a full network capable of learning which micro expressions might give more insight to a person’s emotion. (Shown in Figure 6)

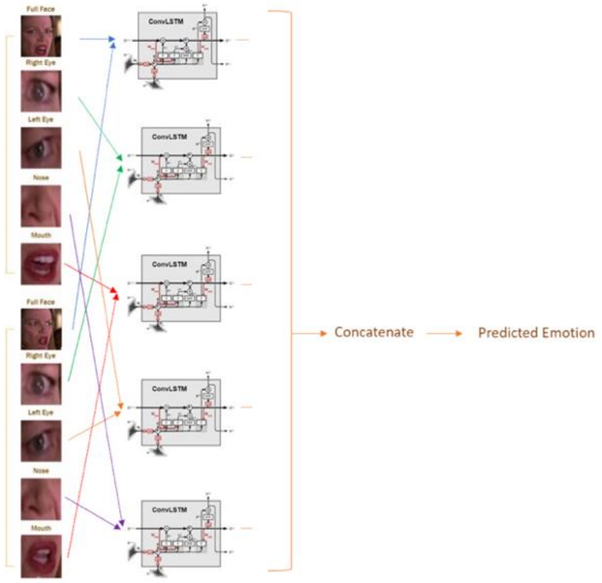


Figure 6: Model Visualisation

E. *Overfitting*

Training a deep neural network such as ours that can generalize well to new data is challenging. To prevent overfitting in our model, we employed a variety of measures. As briefly mentioned in the model, we add a dropout layer after every dense layer. This dropout layer helps to regularize and approximates training our large neural network by randomly ignoring or dropping nodes. Its effect is that after each update to a layer during training, we introduce noise to the data and we now have a “different" configured layer.

Our model also shuffles the training data. Shuffling the data serves the purpose of preventing overfitting by reducing variance. This ensures that our training, testing and validation sets are representative of the overall distribution of the data. This is also computationally more effective when calculating the stochastic gradient descent on a single batch as we will quickly find the true minimum instead of taking small steps every epoch.

Lastly, we also added early stopping to stop our training process as soon as the validation loss rises instead of training for a fixed number of epochs. Not only will this prevent overfitting, but we can also stop the model earlier and save time.

V. Results

To evaluate the accuracy of our framework, we have conducted experiments on the public dataset that we used, BAUM and our own self-created dataset, made up of single face videos with images where the same model was wearing mask and ones without mask. The model is trained with a minibatch size of 8, learning rate of 0.001 and epoch number of 120.

A. *Experimental Results and Analysis*

We applied the same preprocessing process to the test set as we did to our train set and we test the model predictions on this test set that had gone through preprocessing.

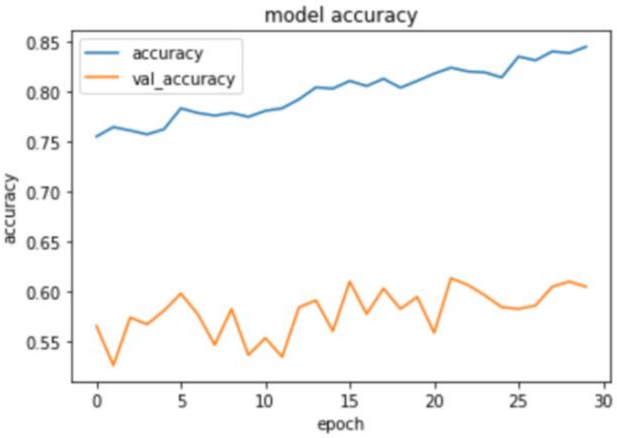


Figure 7: Model Accuracy per epoch

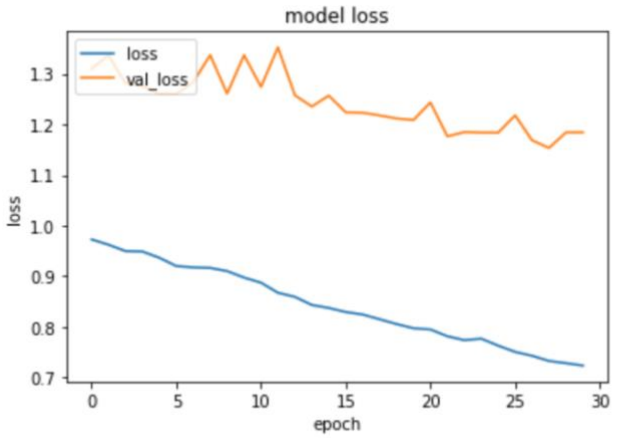


Figure 8:Model loss per epoch

Although we trained the model for 120 epochs, we did it in 4 different runs due to hardware restrictions, saving the model and loading it every 30 epochs. In figure 7 and 8, we can see the plot for the model accuracy and the model loss for the last 30 epochs. As you can see there is a steady increase in the model accuracy and a general decrease in the loss and as it approaches the end of the final set of 30 epochs the values converge. We implemented early stopping while training for this very reason. The model stopped training at epoch number 107. In order to prevent overfitting, we have considered validation accuracy when deciding when to stop training the model.

B. *General Metrics: Model Evaluation on test set*

Table 1

Metric	**Score**
Accuracy	55.20%
Precision	0.527
Recall	0.534
Fscore	0.526

According to these metrics, we can see that the precision, recall and Fscore are all around the same values, this means that the tradeoff between precision and recall is balanced.

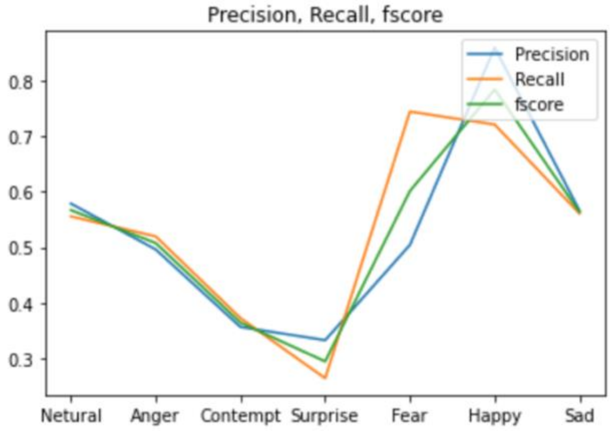


Figure 9:Evaluation metrics for each emotion

As you can see from figure 9, this is a more in depth look at the evaluation metrics where you can see these metrics for each individual emotion. The metrics such as surprise and contempt had lower metric scores while emotions such happy and fear had higher scores. Looking at the significance of this we postulate that happy has such relative high scores possibly because happy is a unique looking emotion compared to the other emotions like contempt. This could also be due to the dataset distribution. There are the most images labelled happy which could be why the scores are so high for it. We see however, that surprise has a very low recall, meaning that it misses out on classifying a high percentage of images that are labelled surprise.

C. *Confusion Matrix*

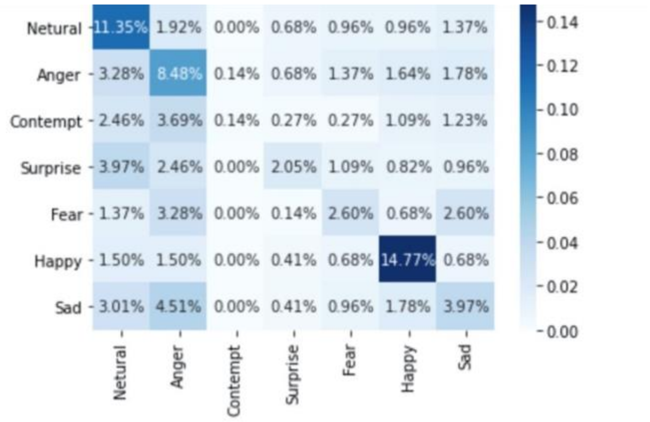


Figure 10: Confusion Matrix for all emotions

Figure 10 shows the confusion matrix of the emotions where each box in the figure represents the probability that emotion y is predicted when the image is that of the emotion x. If the largest probabilities are in the main diagonal of this figure, it represents a good model because this means that a high percentage of the images labelled as emotion x is labelled correctly. The boxes with different background colors are used to indicate the recognized probability. The shade of the boxes indicates the probability of it happening, so the boxes that are darker have a higher probability of happening.

As you can see from figure 10 most of the darker shaded boxes are in the main diagonal and for each label the largest percentage is from the box with the correct corresponding prediction (looking at each column). This indicates that for any test image our model had the highest chance of predicting the correct emotion over some other emotion which is very promising.

Figures 11 to 14 show the confusion matrix for each individual emotion, in each emotion confusion matrix, the top left is the true negative images, which are the images that are labelled as other emotions and predicted as other images. On the top right is the false positive. Where the image is predicted as this emotion, but it's not labelled as this emotion. On the bottom right is false negative images. These are the images that are labelled as that emotion but not predicted as them. Finally, the bottom right is true positive where the images are correctly predicted as this emotion.

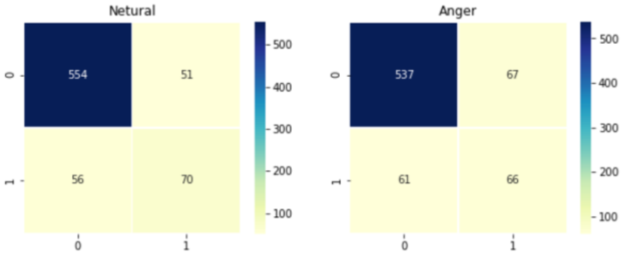


Figure 11: Confusion matrix for Neutral and Anger

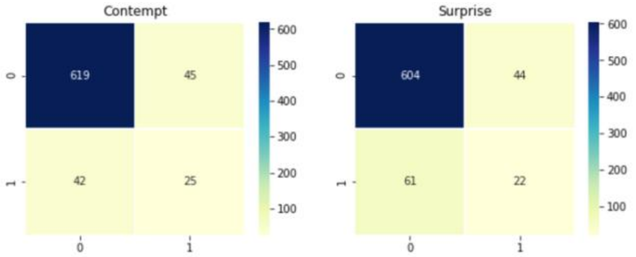


Figure 12: Confusion matrix for contempt and surprise

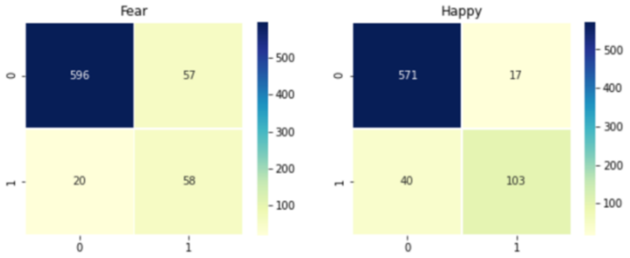


Figure 13: Confusion matrix for fear and happy

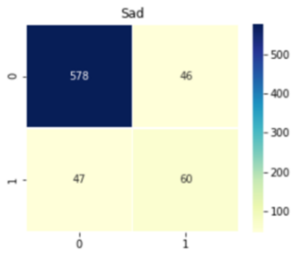


Figure 14: Confusion matrix for sad

Emotions that the models predicted well are like happy and neutral with most of the images being true positive and the emotions that the model is performing poorly on is contempt and surprise, where there are more false negative and false negative images than true positive.

D. Next Highest Emotion

Emotions are not easy to be classified as many of them share many similarities in facial features. Thus, this section is to find out the emotions that our model deems as similar by analyzing the next highest prediction for each image that is predicted with an emotion. For each emotion we have a graph in figure 15 which shows the emotion with the 2nd highest probability to be it. The x value indicates the frame number, and the y value indicates the 2nd most probable emotion. Thus, for the lines that are denser this signals that this emotion is similar to the one for the graph.

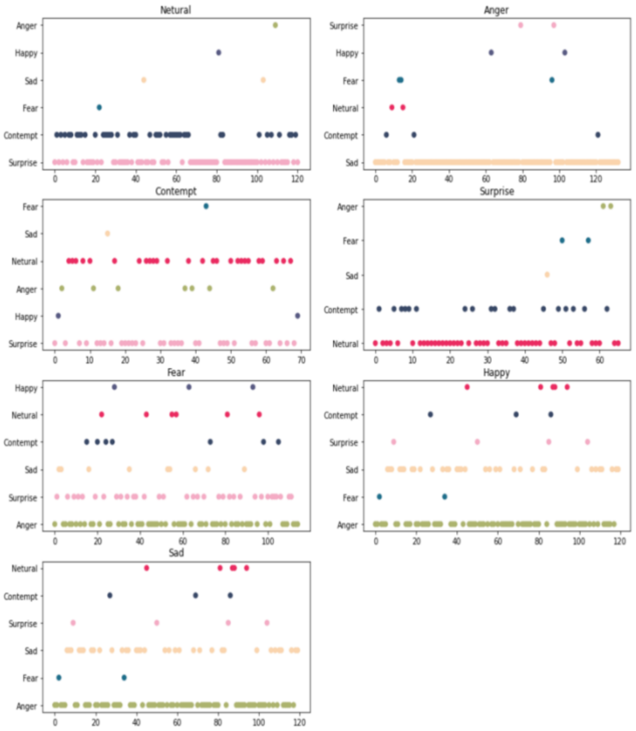


Figure 15: 2nd highest emotion predicted

Some observations we can make from this are that the model views anger and sad similarly and surprisingly anger and happy. We can also see what are very different such as neutral and anger or fear and happy.

E. Comparison with state-of-the-art-models

To have a clearer picture of how our model compares with the models out there. We researched for some models that had a similar approach of video only models that worked on the same Baum dataset which we had used. For both datasets these models only used data from the BAUM-1 dataset. For example, the model created by XianZhang et al. Is a spatial temporal video emotion recognition model. Thus, these models are very similar to ours and we can have a fair comparison of the accuracy of the model.

Table 2

Model	Accuracy
Zhalepour et al.[2]	45.04%
XianZhang Pan et al.[3]	46.51%
MicroX + FullFace	55.20%

The results in table 2 show that we outperform the state-of-the-art strategies on the Baum Dataset by improving the accuracy from 46.51% to 55.20%.

F. Application on in real life dataset

In order to see the effectiveness of our model outside the context of the Baum dataset, we decided to create our own real-life dataset to test the effectiveness of our model in predicting our emotions.

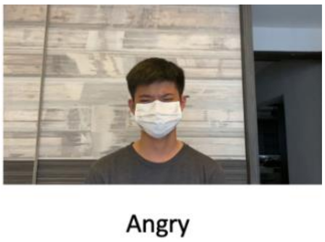


Figure 16: In real life frame with mask

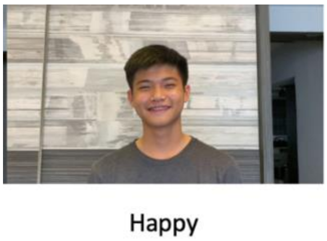


Figure 17: In real life frame w/o mask

As seen from the example frames in figure 16 and 17, we created videos of ourselves in a well-lit environment with one dataset being with mask and one dataset being without. This is to see the ability of our model to still detect the correct emotion without one of the MicroX components. The results for this model can be seen in the table below.

Table 3

Data	With Mask	W/o Mask
Accuracy	17.86%	10.2%

However, these results were not very good. This could be due to the difference in environmental settings with our datasets.

VI. Further discussion and analysis

A. *Greyscale vs Color*

Table 4

Model	Color	Greyscale
Accuracy	55.2%	37.1%

We explored the option of making our frames grayscale instead of color. The rationale for this would be to make the process more computationally efficient and it will take a shorter time to train and more crucially to predict. This might be important in our use case of this model because it can give a more real time response of the emotion on a livestream or a video. This attempt is also done based on our hypothesis that color is not very important for the task of detecting emotion and the pixel intensities provided should be enough to differentiate the facial features from general skin. However, our model seemed to suffer significantly when the input was changed to greyscale. Thus, to maintain accuracy of prediction, we decided to stick with color for higher detection accuracy.

B. *MicroX vs Full Face*

Table 5

Model	Accuracy
Full Face	71.5%
MicroX	23.1%
MicroX + Full Face	55.2%

When creating the model to detect emotion, the first thing we did was to train the model based on only full faces. It reported an accuracy of 71.5%. Following that, we tried our implementation of MicroX as the input training set instead of the full-face frames. Unfortunately, our results in table 5 reflect that MicroX did not work very well compared to taking the full-face frames. After seeing this drastic drop in accuracy, we decided to improve the model by deciding to concatenate another LSTM model with the full face with the other MicroX LSTM models.

The main reason that we postulated on why this might be the case would be that there might be a clear correlation between the different facial parts and the features learned with all parts are more indicative of the emotions than when isolated.

However, we did not just drop MicroX for full face as we still saw some applications where MicroX might perform better. Such scenarios include when the whole face or partial face is blocked, the model will still potentially work as they will leverage on the other parts of the face. This is especially applicable in the current pandemic climate where it is mandatory to wear a mask.

C. *Facial feature most indicative of true emotion*

Table 6

Facial Feature	Accuracy
Left Eye	37.8%
Right Eye	39.8%
Nose	27.6%
Mouth	23.8%

To have a better gauge of which facial parts were the most indicative of the emotion, we did some experiments to find out more about this. We trained the model only on that facial part and the results on the accuracy of predicting the emotion is in table 6. The parts that had the highest accuracy were the eyes and surprisingly the parts with the lowest accuracy were the mouth. Thus, we conclude that the eyes play a large role in the accurate prediction of emotion.

VII. Future Development & Conclusion

In our current model, we simply concatenate our four micro expression and full face. Each micro expression and full face are individually fed into one ConvLSTM block and thus the micro expressions are treated as mutually independent. One possible extension of our model is that we can increase the complexity of our micro expressions network. For example, we might try forming an exclusive network for certain micro expressions that might have more relation to each other such as the left and right eye. We can also try another combination whereby the eyes and eyebrows are extracted together. With different combinations, we can find the optimal network in which our model is able to predict the emotion with higher accuracy. We can also build our model such that several different combinations of micro expression are available. With such highly dense model, even when we feed a video whereby not all micro expression of the person is available, our model can extract all available micro expression and utilize the best available combination to predict the best emotion.

The number of frames per segmentation that is being fed into each LSTM block (currently 10 frames) can be increased to a higher value to aid in and maximize the use of the temporal network. However, this may be a factor limited by the dataset (some having a maximum of 10-20 frames) only. Additional pre-processing could be done to work around this, such as sieving out videos with less than chosen number of frames or even duplicating frames to achieve the desired number of frames required. Although both suggestions might have some downsides to them such as lessening data or force feeding the temporal network altered data, the overall effects may be helpful to the model.

With reference to the results we obtained from our experiments on in real life data in table 3. We feel that we should seek to improve this especially since we want to apply this to potential use cases like identifying driver's state. One way that this can be improved would be to feed new data into the dataset, and this new data should be more representative of the image we would like to predict in our use case such as ones with outdoor lighting.

Most importantly, the use of the eyebrows as a MicroX is a key feature we believe can improve the model even more. Eyebrows are important in informing us more about a person's emotion. We could find a better alternative to retrieving the pixel coordinates of the eyebrows through a different tool for this improvement.

At this juncture, the model has a lot of room for improvement for various implementations, to be trialed with different parameters such as to achieve the optimal results across all variables. Nonetheless, with the current results, we acknowledge that MicroX shows much potential and will be able to make a significant difference in the challenge of emotion recognition.

Acknowledgment

This was supported by

S. Zhalehpour, O. Onder, Z. Akhtar and C. E. Erdem, "BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States", IEEE Trans. on Affective Computing, DOI: 10.1109/TAFFC.2016.2553038.

C. E. Erdem, C. Turan and Z. Aydın, "BAUM-2: A Multilingual Audio-Visual Affective Database", Multimedia Tools and Applications, DOI: 10.1007/s11042-014-1986-2, 2014. Database web site: http://baum2.bahcesehir.edu.tr.

References

[1] **Convolutional experts constrained local model for facial landmark detection** A. Zadeh, T. Baltrušaitis, and Louis-Philippe Morency. *Computer Vision and Pattern Recognition Workshops*, 2017

[2] Zhalehpour, S.; Onder, O.; Akhtar, Z.; Erdem, C.E. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. IEEE Trans. Affect. Comput. 2017, 8, 300–313.

[3] Deep Temporal–Spatial Aggregation for Video-Based Facial Expression Recognition Xianzhang Pan , Wenping Guo , Xiaoying Guo, Wenshu Li, Junjie Xu and Jinzhao Wu, Symmetry 2019