Robots Vs Humans : A comparative study on Video Facial Emotion Recognition task

Kannan Venkataramanan* kv942@nyu.edu

Srishti Grover* sg5783@nyu.edu

Yinzhu Yang* yy3164@nyu.edu

Zhilin Zhang* zz2507@nyu.edu

Abstract

In this project, we have implemented a Convolutional Neural Network based facial emotion recognition (FER) system on basic emotions and conducted a comparative analysis to understand the cognitive processing with which human beings perceive emotions vis-a-vis Video Facial Expression Recognition systems. The experiments have been performed on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, which is a multimodal dataset containing video recordings from actors who express neutral, calm, happy, sad, angry, fearful, surprise, and disgust emotions. We have applied various architectures of CNN (AlexNet, VggNet, ResNet) and their variants to the frames extracted from the videos in order to determine the emotion from the data. The RAVDESS dataset has been classified by 247 participants who were asked to make judgements on three classes: "category of the emotion, strength of the emotion, and genuineness of the emotion". In addition to this, we have collected additional responses from 20 human subjects as well as their reasoning process to validate these results. Finally, we have evaluated the correlation coefficient between classification performance of our FER systems and human judgments to understand and compare the cognitive aspects with which human beings and computers form representations of such tasks.

1 Introduction

In today's world, Human-Machine Interaction systems (HMI) are increasing at an exponential rate and these systems are yet to reach their full emotional and social potential. Facial expression, which

*Equal contribution

plays a vital role in social interaction, is one of the most important nonverbal channels through which HMI systems recognize humans' emotions.

Multiple studies (Bottou and Bousquet, 2008) (Kingma and Ba, 2014) using various different approaches such as Support Vector Machine (SVM) (Kharat and Dudul, 2008), rule-based reasoning (Pantic et al., 2005), genetic algorithm (Rizon et al., 2009), etc., have been conducted to identify and understand the human facial expressions. However, as facial expressions are not simply separate and static images but continuous processes unfold in time, a model using videos as training data could better capture this particular characteristic of human emotional perception. Video Facial Expression Recognition (VFER) plays a vital role in understanding the emotions of human beings and in the enhancement of human interactions with robots and auto-motion mechanisms. In this study, we intend to develop a Convolutional Neural Network (CNN) based facial expression recognition system on basic emotions and conduct a comparative analysis to understand the cognitive processing with which human beings perceive emotions vis-a-vis VFER systems.

As previous literature suggested,(Ilbeygi and Shah-Hosseini, 2012) an emotion recognition system mainly has three processing stages: 1. Face Detection, 2. Feature Extraction and 3. Classification. Here, it is worth mentioning that the second step, feature extraction, is one of the most complicated stages in emotion recognition system. To increase the precision of the results and to streamline the modeling process, it is helpful to extract certain face areas and attributes which are valuable for emotion recognition, so that we could pay extra attention to these features during the modeling process. The features that are used to judge which basic emotion it is in emotional perception could be summarized as follows, ranking by pervasiveness and distinctness according to previous literature : The primary features include the degree of eye opening and the degree of mouth opening, secondary features include the degree of eyebrow constriction and the degree of mouth corners. There are also auxiliary features,including mouth length, nose and nose-side wrinkle, lip thickness, whether teeth exist or not, whether the chin is tightened or loose,though used less frequently. For our current purpose, we mainly rely on primary and secondary features which have proven to be useful in the past.(Chakraborty et al., 2009)

Once the model is developed and validated, we used the model to score the test dataset. At the same time, in addition to the established human data in the RAVDESS, we also conducted behavioral experiment to collect responses from human subjects as well as their reasoning process. The aim of this step is threefold: First, more behavioral data could validate the results.Second, our behavioral data are collected using Chinese subjects, which could add diversity to the previous all-North American subjects pool and make the results more robust.Third, asking human raters to classify the videos and seeking their opinions on why they classified so could provide insights into in what way the models are different from the way human perceive and judge emotions.Finally, we evaluate the correlation

coefficient between classification performance of the FER systems and human judgments [6]. The patterns of performance will be further analysed to understand the cognitive aspects with which human beings and computers form representations of such tasks.

Section-2 discusses the methodology, in which data preparation, collection of emotion categorization data from human subjects, implementation of CNN models is discussed. This is followed by section-3 that discusses the results obtained by the CNN models, visualization of regions of interest for a model for performing classification and Human cognition results. Finally, in section 4, we discuss our findings and conclusions drawn.

2 Methodology

2.1 Data

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is a validatedmulti-modal database of emotional speech and song. This gender-balanced database "consists of 24 professional actors, each performing 104 unique vocalizations with emotions that include: happy, sad,angry, fearful, surprise, disgust, calm, and neutral". There are a total of1440 speech utterances and 1012 song utterances.

2.1.1 Data Preparation

The video dataset of RAVDESS was used for this analysis. First, the input video was decomposedinto frames using opencv package. The data was generated with various frame rates between 0.1seconds to 0.5 seconds. From our preliminary analysis, given the computational capabilities, thefinal modelling dataset was generated with 0.15 seconds as the frame rate which resulted in 48,364images.The videos were of 3 sec length and the first 0.5 sec and last 0.35 sec did not have any emotionexhibited. Hence, this section was culled for the analysis.Emotion recognition experiments are broadly categorized into speaker-dependent and speaker-independent experiments. Actor-dependent experiments contain video instances of different emotionsof the same actor in the the training, validation and test datasets. On the other hand, Actor-independent experiments are experiments where the training, validation and test data consist of audio instancesfrom different actors.In the current experiment, we have generated datasets of both kind. For the actor-dependent analysis,the dataset was split randomly into train,validation and test in the ratio of 80:10:10 and for theactor-independent analysis,videos of actors 1-20, actors 21 22 and actors 23 24 were usedfor training, validation and testing respectively. Since, both validation and testing datasets containinstances of both male (odd actors) and female (even actors) actors, we have ensured that the modelsare not tuned to a particular gender. Table 1 shows the split of dataset amongst train, validationand test for both the actor based and random split

cases. Refer Appendix for complete details.

Dataset Type	Train	Validation	Test
Actor Based	39427	4487	4450
Random Split	38691	4836	4837

Table 1: Train, Validation and Test datasets

2.2 Behavioral Experiment

We collected emotion categorization data from 20 Chinese human subjects to compare the performance of neural networks and real human cognition. Each subject was asked to categorize silent videos in the test set (stimuli of actors 23 and 24) into the eight emotional labels (neutral, calm, happy, sad, angry, fearful, disgusted, surprised) according to their facial expression. They were asked to "choose the one emotion that you think is most correct" while watching each video, and they responded 1-8 for each emotion respectively. Number of watching times was not restricted and participants could watch repeatedly to make the judgments. The video presenting order was randomized. There was no practicing session and participants were expected to make the judgment according to their real-world experience. The overall duration of the whole judgment process was around 25 minutes for each subject. A post-judgment interview was conducted with participants to collect information of their reasoning process for the task.

2.3 Convolutional Neural Networks

For the CNN modeling part, we have implemented AlexNet, ResNet and VGGNet in our project since these deep learning models are well known for classification of images. We have also explored the variants of these models.

AlexNet (Krizhevsky et al., 2012) won the ImageNet 2012 competition and contains about 62 million parameters. Its architecture consists of eight layers: five convolutional layers and three fully-connected layers. Relu activation is applied after every convolutional and linear layer. Dropout is applied before the first and the second fully connected layer.

VGGNet (Simonyan and Zisserman, 2014) scored first place on the image localization task and second place on the image classification task in Imagenet 2014 Competition. VGGNet has multiple variants, in our project, we have implemented VGGNet16. Alexnet has large kernel-sized filters but VGG16 offers an improvement by replacing these with multiple 33 kernel-sized filters one after another. Its architecture consists of 16 layers: 13 convolutional layers and 3 fully-connected layers. All hidden layers have ReLU non-linearity. It accepts RGB input images of size 224 x 224 .

Resnet (He et al., 2016) was introduced in 2015 and made it possible to train up to thousands of layers because of its powerful representational ability. The basic building block of ResNet is a residual block, that consists of "identity shortcut connection". ResNet doesn't have the vanishing gradients problem. Resnet50 is the most common variant of ResNet model and has 48 Convolution layers, 1 MaxPool and 1 Average Pool layer. It has an image input size of 224 x 224. In our project, we have implemented both ResNet50 and its variant SE-ResNet50(Squeeze-and-Excitation Networks ResNets). SE-ResNet50 is formed by adding SE-blocks to ResNet-50 and can achieve almost the same accuracy as ResNet-101, a more deeper version of ResNet than ResNet-50, delivers, hence improving channel inter-dependencies at minimal computational cost.

Additionally, we have also implemented a basic 3 Layer CNN network. We wanted to compare how the predictions are for a parsimonious model as compared to other deep models.

2.4 Transfer Learning

A major advantage of using the selected models is that they can be fine-tuned on any dataset for a particular classification task after loading a pretrained model using transfer learning. For our project, we have applied transfer learning for fine-tuning the models that were pretrained on ImageNet dataset.

All the images are first cropped to size 224 by 224, so that they can be fed to these models. All the models were implemented in python using pytorch library on NYU's Prince Cluster. A learning rate of 0.001 and batch size of 64 were used for training our models for a maximum of 40 epochs. Since its a multi-class classification problem, we have used Cross Entropy Loss function for training our models. Moreover, Stochastic Gradient Descent(SGD) optimizer was used with momentum = 0.9 for updating the model weights.

3 Results

3.1 CNN Model Results

Table 2 shows a summary of best performing models for different model architectures and input features. We can see that the random split models perform much better than actor based models in terms of the accuracies obtained on validation and test datasets. This provides an interesting insight into how a model's learning capacity can get affected by the way data is split across train, test and validation sets. A possible reason behind this may be that randomly splitting data provides more aspects of different actors for the model to learn, but this may not happen in the actor based split. In particular, VGG16 gives the best results across all the models for both random and actor based splits, with Random Split VGG16 performing best giving an accuracy of **97.97%** on test dataset.

Additonally, we also observe that 3 layer CNN model is not performing as well as compared to the deep models and is not able to understand the dynamics of the input image.

Figures 1 and 2 show the loss and accuracy plots for train and validation sets on actor based and random split data. Figure 1a) and 1b) show that, in case of actor based data, the model is not able to generalise well on the test dataset, thus leading to overfitting. But for random split case, the model is able to learn and generalise well on the test dataset, as can be seen from Figure 2a) and 2b).

In an attempt to understand the cases of misclassifications for the best performing VGG16 model, we have depicted the confusion matrices in Figure 3 and 4. We can see that both the random split and actor based cases have some common cases of misclassifications which are - fearful being predicted as sad; happy as calm; fearful as angry. Since these emotions are in realty very similar when expressed by humans, it is difficult even for humans in some cases to correctly guess what emotion a person is expressing, unless some extra context is provided. Hence, we can say that our models are legit and produce logical results.

Model Type	**Architecture**	**Validation Acc.**	**Test Acc.**
Random Split	SE-ResNet50	95%	94%
Random Split	ResNet50	97%	95.5%
Random Split	AlexNet	97.2%	97.02%
Random Split	VGG16	98.7%	97.97%
Actor Based	SE-ResNet50	53%	52.5%
Actor Based	ResNet50	49.8%	48%
Actor Based	3-Layer CNN	48%	46%
Actor Based	AlexNet	55.8%	49.89%
Actor Based	VGG16	58.7%	53.01%

Table 2: Accuracy results for the CNN models

3.2 Human Behavioral Results

This experiment was performed amongst all the human participants to understand the differences in judging an emotion by different humans. In the next section, we shall compare the emotion recognition between models and humans.

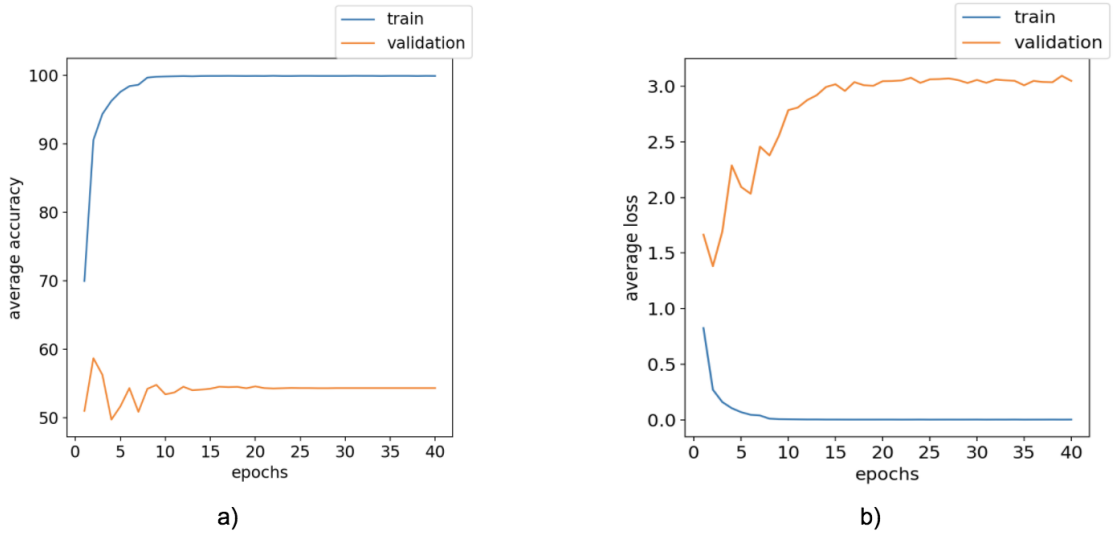


Figure 1: Accuracy and loss plots for VGG on actor based data

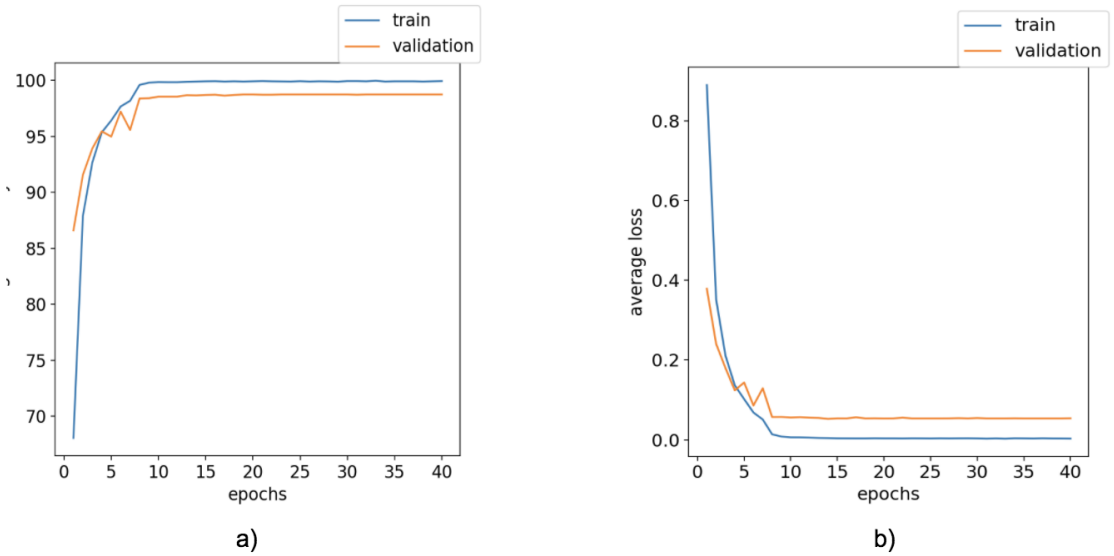


Figure 2: Accuracy and loss plots for VGG on random split data

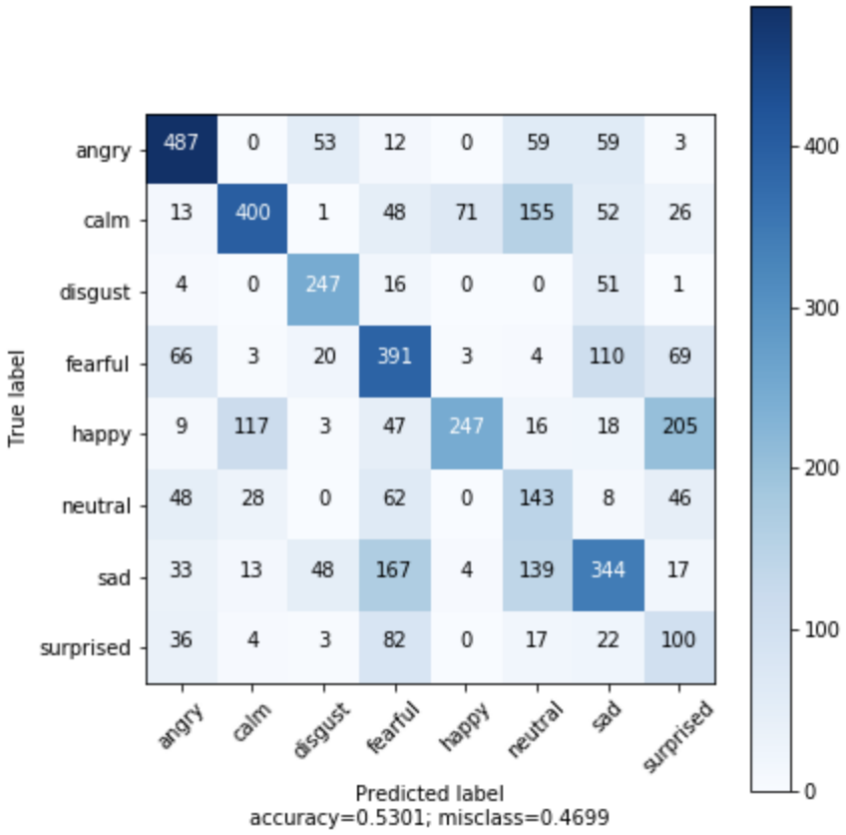


Figure 3: Confusion Matrix for VGG on actor based data

The mean accuracy among all participants and all stimulus conditions was **62.60%**, with individual accuracy ranging from 47.12% to 71.63%. Most participants' accuracy falls between 60% and 70%. Correlation coefficients between the correct answer and each subject's judgments were also calculated as indicator of within-group consistency and a supplement to accuracy. Lowest value was 0.3622 and highest was 0.8433, with an average of **0.6387**.

Figure 5 shows the response distribution among different emotion types in terms of percentage of choosing each emotion type when an intended emotion is presented, and each column adds up to 100%. Judgement performance was good with Happy, Angry and Disgust emotion types, with an accuracy above 80%. Distinction between Neutral and Calm was not very clear in human judgment, as is in the data validation results from the database. Judgment for Calm and Surprise categories was largely distributed among different other emotion types. In addition, participants were prone to judging Calm as Happy, Happy as Surprise, Sad as Neutral or Calm, Angry as Disgust, Fearful as Surprise, and Disgust as Sad.

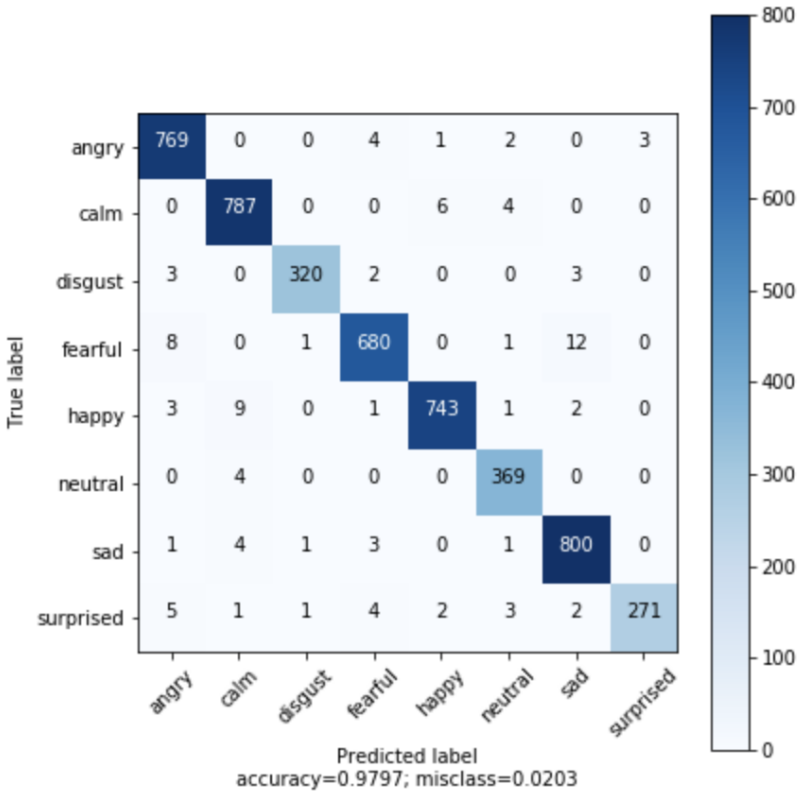


Figure 4: Confusion Matrix for VGG on Random split data

		Actor intended emotion							
		Neutral	**Calm**	**Happy**	**Sad**	**Angry**	**Fearful**	**Disgust**	**Surprise**
Subject chosen emotion	**Neutral**	40.94	18.75	2.34	12.97	1.41	2.50	0.94	15.94
	Calm	41.56	38.91	1.72	13.44	0.94	0.78	0	6.56
	Happy	5.00	35.78	78.75	1.25	0.31	0.31	0	10.00
	Sad	1.56	1.25	0.47	59.84	0.31	1.41	12.50	1.88
	Angry	8.44	2.34	0.16	5.31	85.31	6.72	3.75	14.06
	Fearful	0.94	0.31	0.47	4.69	0.63	67.34	1.56	4.69
	Disgust	0.94	2.50	0.16	1.72	10.63	8.75	80.31	14.69
	Surprise	0.63	0.16	15.78	0.78	0.47	12.19	0.94	32.19

Figure 5: Subject Responses

3.3 Comparison between Model and Human

To compare human behavior and model performance, we analyzed the output of the VGG16 network with best performance as is shown above. Frame-wise judgment results from the model were merged to form the model judgment of each individual test video clip. The mode among several frame results were taken as the result of the video. For the very few vacant data points the average among all outputs were fitted in for data processing concern. Overall judging accuracy of random split model was **84.13%** and accuracy of actor-wise model was **62.50%**. Correlation coefficients between model outputs and correct label was 0.5593 for the actor based VGG16 model, and 0.8528 for the random split model, indicating a reasonable correcation with the actual correct answer in the data set labels. Response distribution of the two models are shown in Figures 6 and 7, with percentage of choosing each emotion type displayed in the grids.

Compared to results with human subjects, the better-performed random split model had an apparently more centered response distribution with higher rate of choosing the right answer and much fewer alternative options, while the actor-based model output were relatively more distributed but still with fewer alternatives than human judgments. In general, random split model showed a more similar pattern to human cognition, with higher performance for Happy and a relatively low accuracy for Surprise. In terms of mean correlation coefficient between model output and human judgment of 20 subjects, random split model is more similar to human behavior with a correlation coefficient at **0.6183** while actor based model only has **0.5044**.

		Actor intended emotion							
		Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprise
Random Split Model chosen emotion	Neutral	87.50	0	0	3.13	0	0	0	0
	Calm	6.25	93.75	0	0	0	0	0	0
	Happy	0	0	87.50	0	0	0	0	6.25
	Sad	6.25	6.25	12.50	96.88	12.50	0	37.50	50
	Angry	0	0	0	0	87.50	6.25	0	0
	Fearful	0	0	0	0	0	93.75	0	6.25
	Disgust	0	0	0	0	0	0	62.50	0
	Surprise	0	0	0	0	0	0	0	37.50

Figure 6: Random Split Model Output

3.4 GradCAM

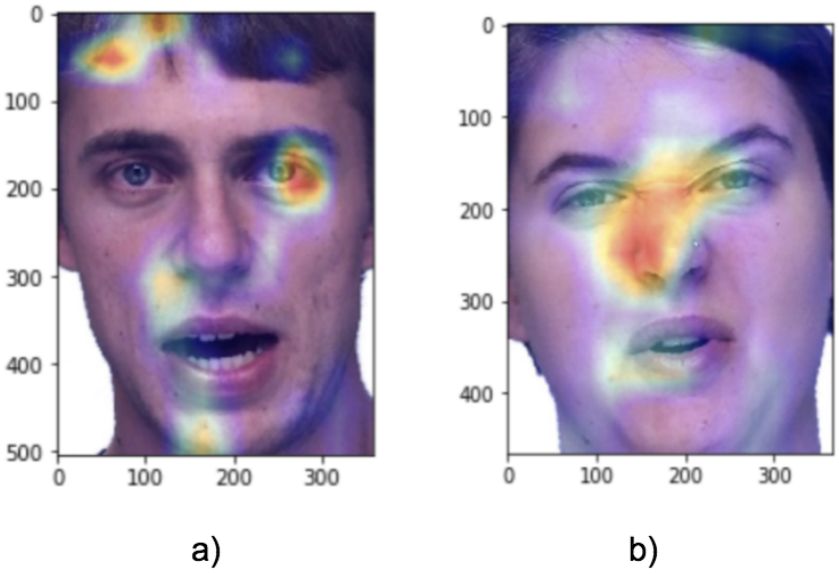
In this section, we have tried to explore the regions of the human face that are important for the model to classify it into the eight emotions. In order to give a visual explanation of how the learning is happening, we selected our best performing model and then chose 4 test sample frames. We

		Actor intended emotion							
		Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprise
Actor Based Model chosen emotion	Neutral	50	18.75	0	28.13	0	3.13	0	0
	Calm	0	62.5	18.75	0	0	0	0	0
	Happy	0	6.25	37.50	0	0	0	0	0
	Sad	0	0	0	37.50	0	9.38	0	0
	Angry	12.5	3.13	0	3.13	96.88	9.38	0	12.50
	Fearful	25	6.25	3.13	28.13	0	68.75	0	31.25
	Disgust	0	0	0	3.13	3.13	0	100	0
	Surprise	12.5	3.13	40.63	0	0	9.38	0	56.25

Figure 7: Actor Based Model Output

then applied GradCAM, a technique for visualizing the regions of input that are "important" for predictions of CNN based models. GradCAM uses the gradient information flowing into the last convolutional layer of the model to understand each neuron for a decision of interest. A heatmap of regions of interest is created and then superimposed onto the original image

Figure 8 show the results of GradCAM implemented on the selected test samples. We can observe that the model places more importance on regions near eyes, nose, lips for its decisions. We also observe that although the model is able to correctly guess the emotion of disgust, however it gets confused between happy and angry emotion, thus sometimes producing false positives as suggested by Figure 8d). However, on close consideration of Figure 16, one may not be wrong in believing that even humans might go wrong in deciding the emotion suggested by the picture since it equally resembles happy and angry emotions.



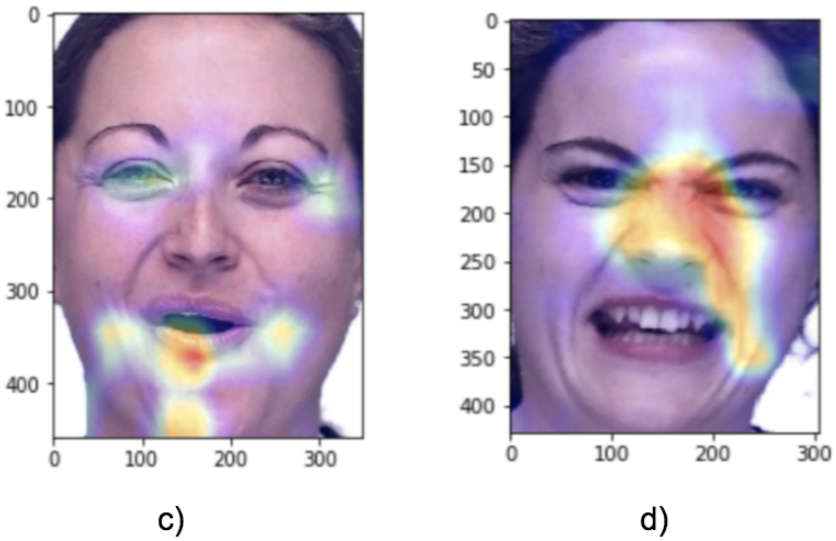


Figure 8: GradCAM results on 4 test samples a) (angry, angry) b) (disgust, disgust) c) (happy, happy) d) (happy, angry) Here for each pair of labels, the first label is the actual class and second is the predicted class.

4 Discussion

In this project, we have tried to perform a comparative study of human emotion recognition between CNN based models and humans, in an attempt to investigate and understand how the learning of CNN based models differ from that done by humans.

For the model learning, we implemented multiple CNN based models and their variants to explore and find the best performing model to serve as a reference for comparison with the human cognition part of this project. The best performing model was VGG16, that achieved 97.97% test accuracy for the random split data. Moreover, the results obtained with the random split based models were observed to be better than actor based models. Also, for understanding the important regions of interest in an actor's face for a model, we implemented GradCAM using VGG16 on four selected test samples.

Compared to the logic of neural network, human participants in this study tend to use a much more holistic instead of analytical way to judge emotion out of facial expression. Many participants reported that they made the judgment out of an "overall impression" rather than detailed, analytical observation of each part of the face. This implies that part of the visual emotional judgment processing which is more similar to our neural networks may happen at a more implicit level where detailed analysis may hide beneath consciousness. However, this gestalt processing does not mean that human beings are not utilizing partial details to evaluate the possible emotion. As previous literature suggested, many participants were paying attention to the extent of eyes opening and mouth opening to help them

make the judgment. Specifically in a dynamic visual presenting form as video, they also used head movement as a supplementary clue to tell the emotion.

Previous study (Livingstone and Russo, 2018) has validated the labeling results in the data set by human participants, with an approximate accuracy at 70%. In our study, human judgment accuracy is lower than the data validation results for two possible reasons. Firstly, our participants were presented with the video-only stimuli, which prevents them from using hints from audio modality. Secondly, the participants in our study were all Chinese but the actors are non-Asian, which leads to a possible cultural difference between our subject population and previous subject population in dataset validation.

Our current study is based on the working assumption that perception of the basic emotions relies on the recognition of facial expressions measured by detecting certain features of certain face areas. However, the emotional perception process of human beings is more complicated than that. Psychologists and neuroscientists (Barrett, 2017) have argued that emotion perception and classification is not merely about image recognition, but should be viewed as a construction which would be also contributed by interoceptive inputs and cognitive interpretations.

In the former case, some of our subjects have reported that they would consciously or unconsciously use their muscles to mimic the expressions in the videos when categorizing. This is also consistent with the finding of mirroring effect, and the finding that when injected with Botulinum toxin, a toxin which paralyzes facial muscles and makes it difficult to make expressions, people show significantly lessened ability in identifying other peoples emotions.(Neal and Chartrand, 2011)

In the latter case, and more generally, it is well established that there are context effects during emotion perception.(Barrett et al., 2011) That is, other than bodily reactions and interoceptive awareness, visual scenes, other faces, words, voice and cultural orientations could all shape the perception of emotions in human faces. This implicates that human beings are not simply reading emotions from a face like reading words from the computer screen. Instead, they incorporate contextual clues when making inferences. A simple example is that when seeing a shouting woman's face with eyes closing, mouth opening with bared teeth, face tightening and jaw clenching, you might judge that she is angry or in pain, but when the whole figure is revealed and you see her wearing sportswear with a bat in her hand, fist clenching and there are people clapping their hands, you would then confidently conclude that she is ecstatic, possibly from winning a game. One important factor among the contexual features that worth specific mentioning is cultural context. It has been shown that cultures differ in the precise facial actions used to pose discrete emotion categories. For example,when looking at startled and sneering faces, Western Caucasian perceivers fixate around the eyes, nose, and mouth of a target face, whereas those from an East Asian cultural context fixate primarily on the eyes. The results of our

behavioral studies have also shown the difference between different cultural populations.(Jack et al., 2009)

For the abovementioned limitations, there are a few possible directions. Though it is hard to develop an embodied recognition system incorporating bodily movement, a doable solution to this problem might be using clues from the auditory modality. In the previous literature mentioned above, subjects without facial muscle movements are tested on visual-only stimuli. If they are presented with auditory stimuli simultaneously, then they could using both clues from the environment and use multisensory integration to adjust their judgements nonetheless, thus reducing the influence of not being able to use interoceptive information. One promising future direction, then, is to combine the audios from the RAVDESS dataset to further generalize our current finding. For the cultural context problem, a possible solution might be to develop different models tailor-made to each population so that the trained models would be separated into several subparts. Another possible solution is to select subjects as a mix of population from different cultures and see whether their difference could be remedied or partially remedied when combining the audio data and some more generalized features could be extracted.

Another possible direction to combine our modeling methods with human cognition is to compare the heat map generated from the models to human eye-tracking data. In this way we can obtain a more thorough perspective into the comparison between saccadic movement and machine recognition. In addition, since we are dealing with videos, a dynamic type of stimuli, the time course with which facial expression recognition can also be explored with eye tracking technology. A more detailed investigation of human eye movement in facial expression, compared with the current literature, is needed to go deeper into the human-machine complementarity.

References

L. F. Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

A. Chakraborty, A. Konar, U. K. Chakraborty, and A. Chatterjee. Emotion recognition from facial expressions and its control using fuzzy logic. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 39(4):726–743, 2009.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

M. Ilbeygi and H. Shah-Hosseini. A novel fuzzy facial expression recognition system based on facial feature extraction from color face images. *Engineering Applications of Artificial Intelligence*, 25 (1):130–146, 2012.

R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, and R. Caldara. Cultural confusions show that facial expressions are not universal. *Current biology*, 19(18):1543–1548, 2009.

G. Kharat and S. Dudul. Human emotion recognition system using optimally designed svm with different facial feature extraction techniques. *WSEAS Transactions on Computers*, 7(6):650–659, 2008.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5), 2018.

D. T. Neal and T. L. Chartrand. Embodied emotion perception: amplifying and dampening facial feedback modulates emotion perception accuracy. *Social Psychological and Personality Science*, 2 (6):673–678, 2011.

M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005.

M. Rizon, D. Hazry, M. Karthigayan, R. Nagarajan, N. Alajlan, Y. Sazali, J. Azmi, and R. Suryani. Personalized human emotion classification using genetic algorithm. In *2009 Second International Conference in Visualisation*, pages 224–228. IEEE, 2009.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

A Appendix

A.1 Materials

The RAVDESS dataset is very rich in nature given that it does not suffer from gender bias, consists of wide range of emotions and at different level of emotional intensity. Other datasets, such as Surrey Audio-Visual Expressed Emotion (SAVEE) and Toronto Emotional Speech Set (TESS), consisted of audios from only male and female actors respectively. We also observe that the RAVDESS dataset is equally distributed across all emotion classes ($\sim$ 15%), so it does not suffer from any class-imbalance problems.Additionally, extensive validation and reliability tests have been performed by the creators of RAVDESS dataset. From a "pseudo-randomly chosen set of 298 stimuli, consisting of 174 speech and 124 song presentations," 247 naive participants were asked to make three judgements on three classes: "category of the emotion, strength of the emotion, and genuineness of the emotion" (**?**, p. 12). From the above figure, we observe that approximately 73% of the rater chosen emotion were well-acted by the actors, ensuring the reliability of the classification of the emotions and the audio content.

A.1.1 Data split amongst classes

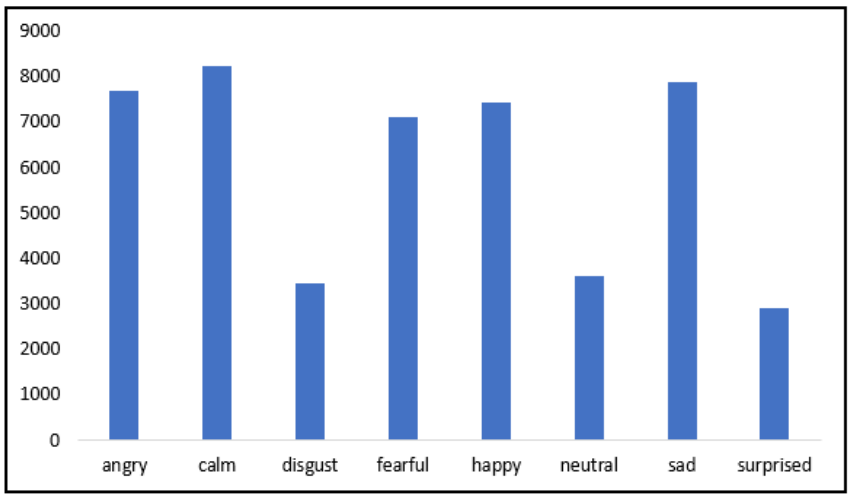


Figure 9: Distribution of images across Emotions

A.1.2 Face Detection

Haar Cascade Face Detector and Deep learning based Face detection algorithms in OpenCv were considered in this experiment. Haar feature based face detection consists of various filters(shown below) to detect if a small section of image can be classified as a part of the face or not. OpenCV's

DNN based face detector is Single-Shot-Multibox detector which uses ResNet-10 Architecture as backbone to predict if a particular image is a human face or not. From our analysis, we observed that DNN based model performed better in not only predicting the face accurately but also in discarding unnecessary background information. Hence, all images were further processed to extract only the face and discard other information.

A.1.3 Data Pre-processing

All the frames were first cropped to size 224 by 224, and then normalised based on the statistics of the dataset. They were then fed to the models.

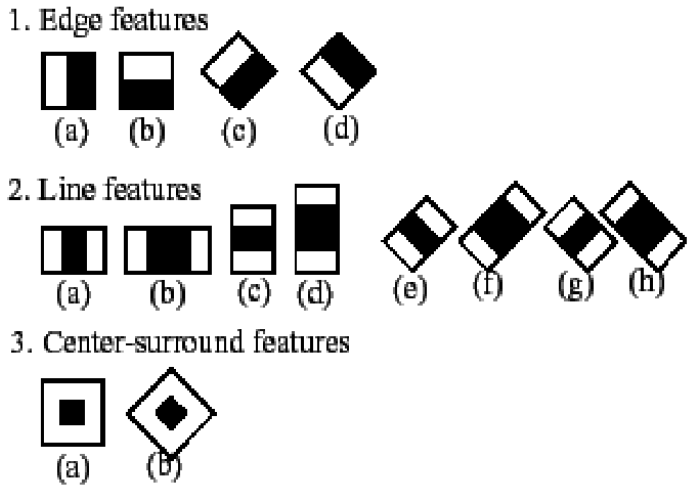


Figure 10: Haar Cascase Face Detector Filters

A.1.4 Face Alignment

Face alignment is another pre-processing step in this experiment. The actors in the video have some gestures when talking or singing, which will cause an error during prediction. Hence, standard face image rotation technique has been implemented.The following steps were carried out to estimate the angle of rotation

- Identify the location of eyes using the detector mentioned in previous step
- Compute the middle point of both the eyes
- Connect both these points and identify the angle with the vertical line

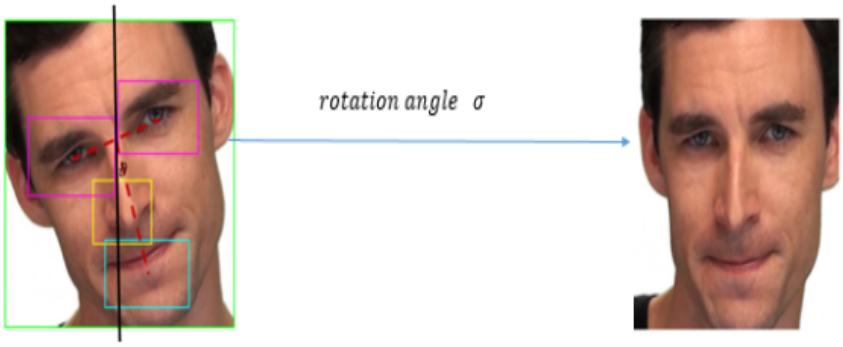


Figure 11: Image Rotation to Align Faces