Facial Emotion Detection Using CNN (Convolutional Neural Network)

Akshat Pareta

Chandigarh University

20BCS6567@cuchd.in

# Abstract

*Numerous intriguing studies* *on the automatic detection of facial emotions have already been carried out recently (FED). Human centre computing and the most recent developments in emotional artificial intelligence are two areas where FED has been applied to improve human-machine interactions (EAI). Researchers working in the subject of EAI want to improve how adept computers are at spotting patterns in human behaviour and facial expressions. The largest impact on this area has come from deep learning since alternative designs are being created to tackle ever-harder issues as a result of the considerable evolution of neural networks in recent years. This article will discuss the most recent advances on automatic expression recognition in relation to computational intelligence using the most recent deep learning models. We show how model which make use of architecture-related methods, including databases, and FED that is based on deep learning can work well together to produce results that are extremely accurate.*

# Introduction

It is easier to identify facial expressions on a person because of their major and distinctive features. FED is a change in face expression’s caused by an person's emotive state on the inside. In addition to machine learning, image analysis, and artificial intelligence, It is utilised in a variety of human-computer interface (HCI) applications, including the analysis of facial images, the recording of people's faces in surveillance, and facial animations.The challenging subject of automatic facial expression identification has drawn the attention of numerous academics recently. The feature extraction stage is crucial in FED. In the literature, According to Alek et al.[1] , oral and written communications represent 38 percent and 7 percent, respectively, to overall transmission, whereas facial emotions account for 55 percent of it.

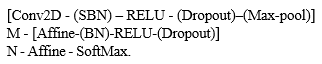
A FED system can be designed using one of two main methods. Some methods start with a succession of images that range from a neutrality face to the strongest emotion. As a result of the limited data they can access, other systems, contrasted with, only employ a single image of individual’s face to detect ’s exactly, and thus frequently perform worse than leading methods [2,3]. A single or both of those characteristic groups are utilised by a FED system., and in addition to the approach type it models, features used in the recognition process are also categorized in this manner. The position of The face organ and skin texture provide the initial set of features. The second sort of characteristic is indeed a geometrical element, that is used to assess a single frame or even a series of pictures through providing details regarding distinctive facial positions & locations. observing how well the locations and points change over time. Face landmarks can be utilized as a starting place for geometrical feature extraction in one technique. During a facial analysis, landmarks are significant areas of the face that can be utilized to collect information. Numerous studies on the subject of recognising facial landmarks have been done; nevertheless, they are not pertinent to our work.

1. **Related Work**

There have been significant advances in the development of automatic expression classifiers. in recent years [7, 8, 9]. Some systems for recognizing facial expressions classify the expression into a range of fundamental emotionss., including joy, sad and rage.. In an effort to provide an unbiased description of the face, others have made an effort to identify the specific muscle movements that the face is capable of doing[11]. The most widely used psychological framework is the Face Action Coding System (FACS)[12]. for summarizing practically all facial movements. Using Action Units, the FACS system classifies human facial movements according to how they appear on the face (AU). A facial expression usually originates from one of the 46 atomic units (AUs) of face movement or related deformation that can be detected. Several AUs are usually added together to form an expression [7, 8]. Additionally, Multilevel Hidden Markov Model, Neural Networks, and Bayesian Networks (HMM) have all seen development in the approaches employed for face emotion identification [13],[14]. Several of them have problems with timing or detection rates. combining two or more techniques to accomplish precise recognition allows for the extraction of features as necessary. Because of illumination and feature extraction, Each technique's effectiveness is reliant on image pre-processing.

# Methods

To assess how well these models performed at identifying facial expressions, we created CNNs with varying depths. For our analysis, we took into account the following network design.



All of these layers have the convolution layer and ReLU non - linear., are referred to as the first component of the network. These layers also include dropout, max-pooling, and spatial batch normalization (SBN), which can be present. After M convolutional layers, which are always Affine-operating & ReLU non-linear and may additionally the network is when batches normalizing (BN) as well as dropouts led to N fully linked layers. In the show's processing order, the affine layer follows the network and computes the score and softmax loss function.. The user has control over the created model's convolution and fully - connected layers layer counts along with the presence of batch normalisation, dropout, and max-pooling layers. We used L2 regularisation in addition to dropouts and batch normalisation methods. Additionally, the user can specify the quantity of filters, strides, and zero-padding; otherwise, basic settings are used.

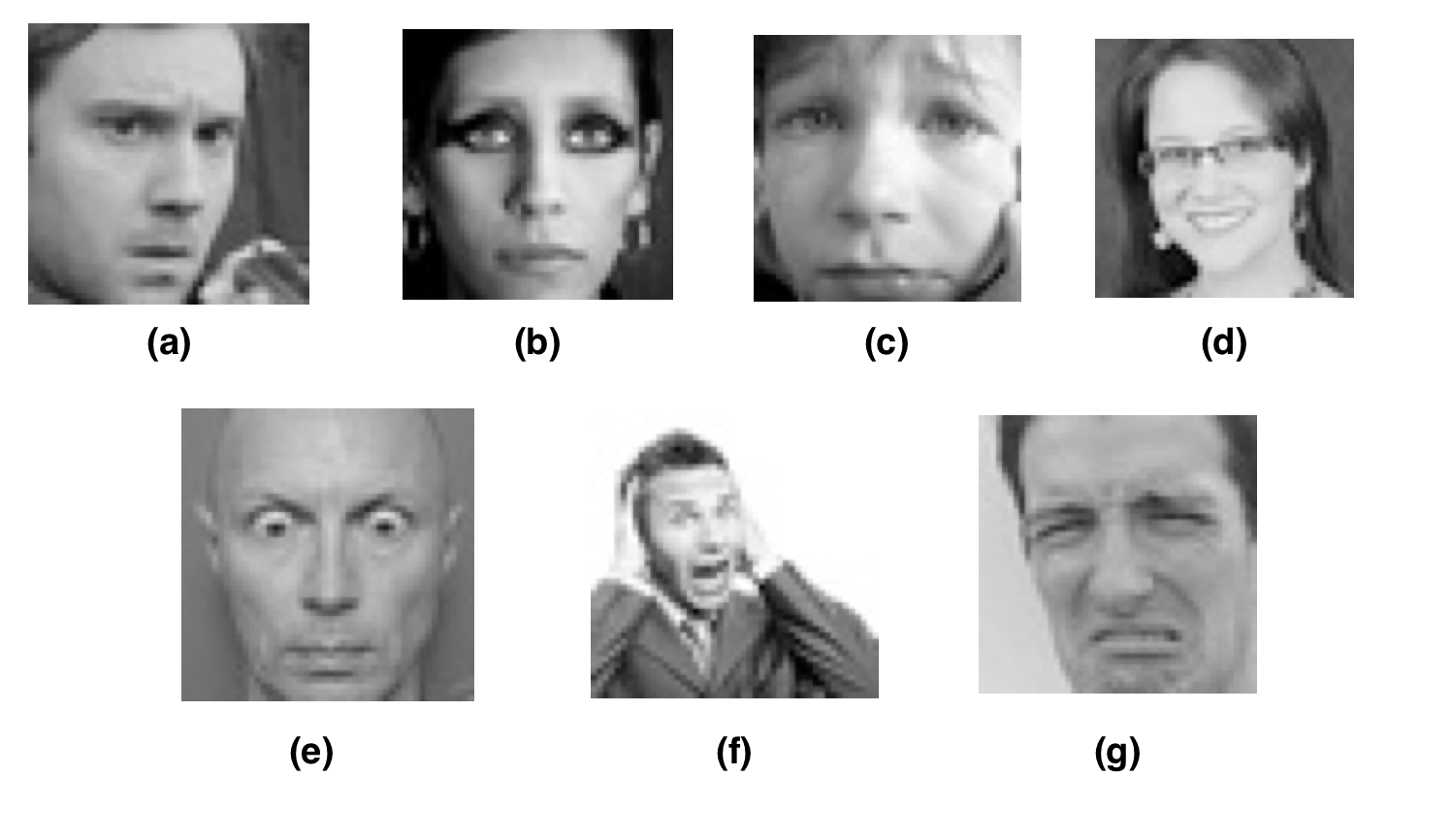
We introduced the idea of merging HOG characteristics with those recovered by convolutional layers using raw pixel data, as we shall explain in the following section. To do this, we used this framework as before, but we now added the HOG characteristics to the features exiting  final convolutional layer. The hybrid feature set is then passed to the fully connected layers, who use it to compute scoring and losses. 

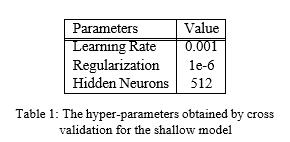
Figure 1: Examples of the 7 face expressions we are taking into account for this classification issue. Aside from being furious[a], you can also feel neutral[b], sadness [c], delighted[d], surprised [e], fearful[f], or disgusted[g].

Thus order to speed up the network train program, we developed that above models in Torch and used GPU acceleration deep learning characteristics.

# Dataset and Feature

For this study, we was using a data from the Kaggle website, which consists of approximately 35000 well-structured 48x48 pixel grayscale photos of faces. Each face in each image about equals the amount of space it takes up in the processed photographs, which virtually perfectly centre the faces. The seven classes that represent various facial emotions must be applied to each image. These facial expressions have been divide in 6 classes: (i) represents rage,(ii) disgust, (iii) fearfull, (iv) happness & joy (v) sadness or grief , (vi) surprise, and 6 neutral. Each category of facial expression is represented by one example in Figure 1. The provided photos were divided into three different sets known as train, validate, and testing set, respectively. along with the images classes no. (between zero to six). In order to normalise the raw pixel data after reading it, we subtracted the average of each image's training pictures, including those taken from the validation and testing set.. We created mirrored images by horizontally flipping photographs from the set is used to train for the enhancement of data.

The characteristics produced by convolution layers using the raw image data were mostly used to classify the expressions. As an additional experiment, we created learning models that fed the input features into Fully Connected (FC) layers by concatenating the HOG features with those produced by convolutional layers.



# Analysis

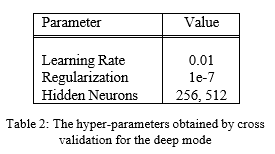
## Experiment

We first constructed a shallow CNN with the help of this project. One FC layer and two convolutional layers made up this network. 32 3x3 The first convolutional layer contained filtering with such a single-length stride with batch normalisation, dropouts, and also no maximum pooling. In the second convolution layers, 64 3x3 filter with such a strideod size of 1 were combined with batch normalisation, dropout, and max-pooling using 2x2 filters. In the FC layer, we employed 512 neurons inside a hidden units using Softmax as that of the losses formula. Rectified-linear Unit (ReLU) also was utilised as just a layer-specific activation function. To ensure that the network's implementation was accurate before training our model, we performed a few sanity checks. The initial loss when regularization is not present was calculated as the first sanity check.

Since there are 7 different classes in our classifier, we anticipated receiving a result around 1.95. A tiny subset of the training data was used to attempt an over-fit of our model as a second sanity check. Both sanity tests on our shallow model were successful. Following that, we began completely training our model. We made use of Torch's Deep learning capabilities that are GPU-accelerated to hasten model training.

To study the effects of adding convolution layer & A stronger C-NN was trained using two FC levels and four convolutional layers.. There were 64 3x3 filters in the first convolutional layer, 128 5x5 filters in the second, Third layer: 512 3x3 filters; fourth layer: 512 3x3 filters. All of the convolutional layers in our model have ReLU as that of the activation function, along with a stride of length 1, batch normalisation, dropout, and max-pooling. The buried layer inside the initial FC layers only had 256 neurons, but the next Fully - connected layer had 512 neurons. Batch normalisation, dropout, and ReLU are similar to the convolutional layers. were utilized in both FC layers. Additionally, Softmax served as a result of our loss. Figure 2 depicts the design of this convolutional model. We performed preliminary loss checking before training the network, as we did with the shallow model, and we looked at the possibility the risk, when only a tiny portion of the training data is used, of overfitting the network. These sanity checks revealed the network implementation to be accurate. Then, using each of the pictures from of the training set, the network trained with 50 epochs and a batch size of 128..In order to obtain the model with the maximum accuracy, we also cross-validated the hyper-parameters.

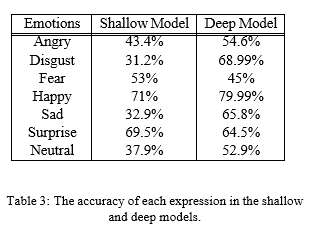
Moreover, we developed networks with five and six convolution layer.to investigate the deeper CNNs, but these networks did not improve classification accuracy. the model has two FC layers and four convolutions layer was therefore deemed to be the best network for our dataset. We solely utilized the features produced by the convolution layers in both the shallow and deep models, the important components of our classification task are taken from the original image pixels. Because HOG characteristics are sensitive to edges, they are frequently employed for facial expression identification. In order to see how well the model performs when it combines two separate characteristics, we decided to see if there was a method to Raw pixels and HOG characteristics should be added to our network. To do In order to accomplish this, we created an unique training algorithm composed of two neural nets, which first contained convolution layer as well as the second of which only contained fully-connected-layers. The first network's features are combined with the HOG features to create hybrid features, which are then sent into the second network. In order to assess the hybrid network's performance, We created two networks by training them similar to the deep and shallow networks we developed in the earlier experiment. In this case, the shallow model's accuracy was rather equal to a shallow model that solely employed raw pixels. The deep model's accuracy was comparable to what we were able to achieve to our own deep model, which utilised raw pixel attributes.



## 5.2. Outcomes

We charted lost history and measured these models' precision. to assess how well the shallow model and deep model performed. These outcomes are shown in Figures 3 and 4. The validation accuracy was increased by 18.46% thanks to the deep net effort, as shown in Figure 4Along with L2 regularisation, additional non-linearity and hierarchy anti-overfitting techniques, such as batch normalisation and dropout,may be seen in the deep network, which has also been seen to lessen the overfitting behavior of the learning model. The training accuracy quickly reached its greatest value, as shown in Figure 3, and the shallow network converged more quickly.

The confusion matrices for the deep and shallow networks were also computed. A depiction of a confusion matrice is shown in Fig 5 and 6. These numbers show that for the majority of the labels, the deep network produces higher correct predictions. It's intriguing to observe how well both models predicted the happy label, suggesting that it may be simpler to acquire the characteristics a smile rather than other emotions.. These matrices also show the labels that the trained networks are most likely to misinterpret. As an illustration, we can observe the relationship between the furious label and the fear and sad labels. There are many situations where people are misclassified as fearful or depressed when their genuine emotion is anger. These errors match what we notice when viewing photos in



the data; Even for humans, determining whether a face is sad or angry can be difficult. This is because not everyone expresses their emotions in the same manner.

We evaluated the efficiency of every models for every expression in additional to confusion matrices. This data is shown in [3]. The In both deep and shallow models, the accuracy in detecting The highest worth of all of the sentiments is a happy expression., as can be seen in the table. Deep network technology has also improved categorization accuracy for the majority of phrases. In other words, adding more features does not always result in superior features for particular expressions.

As stated in the previous section, we created mixing characteristics generated by convolution layer and learning models with HOG features

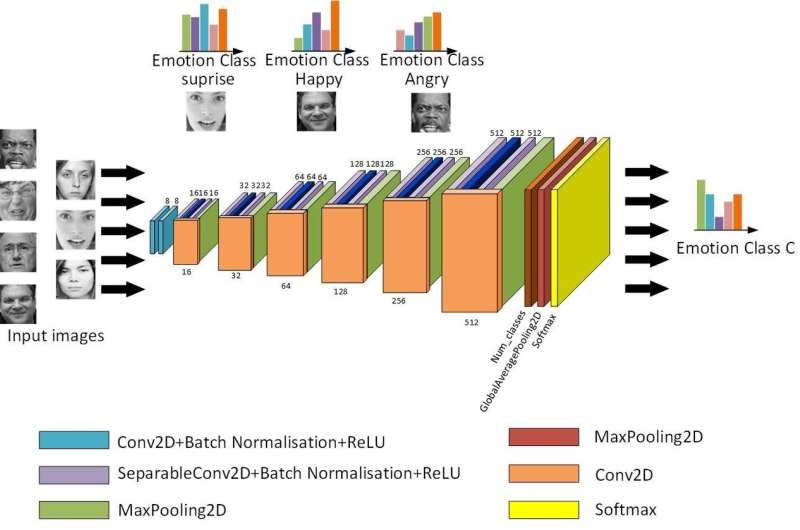
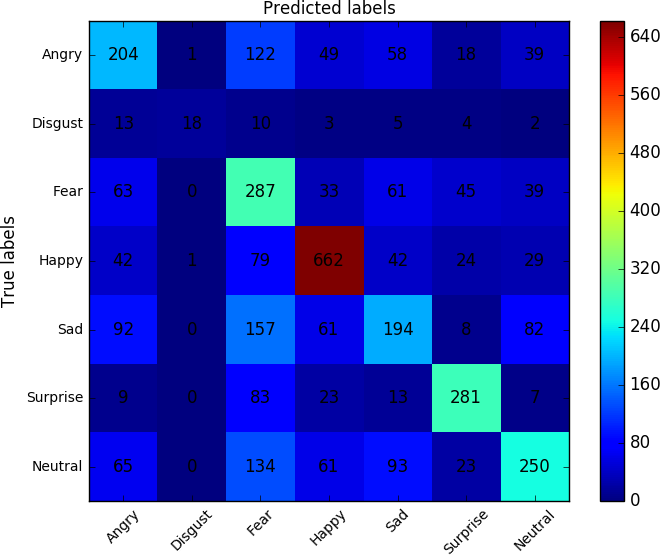
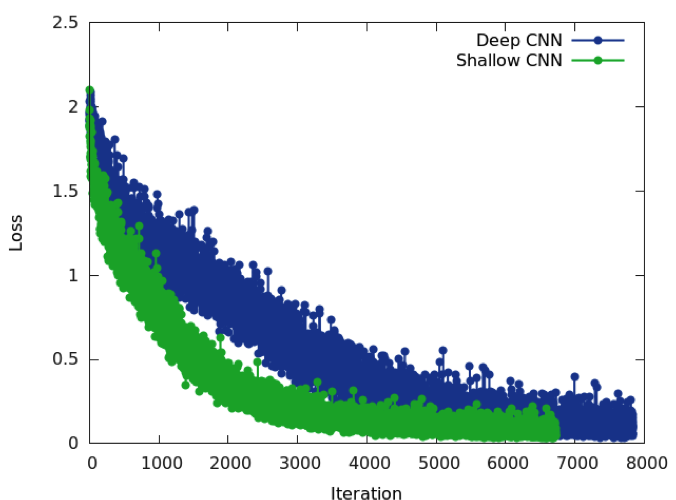


Fig 2: The deep network's design consists of two fully linked layers and four convolutional layers.



Fig 3: The deep and shallow models' loss histories

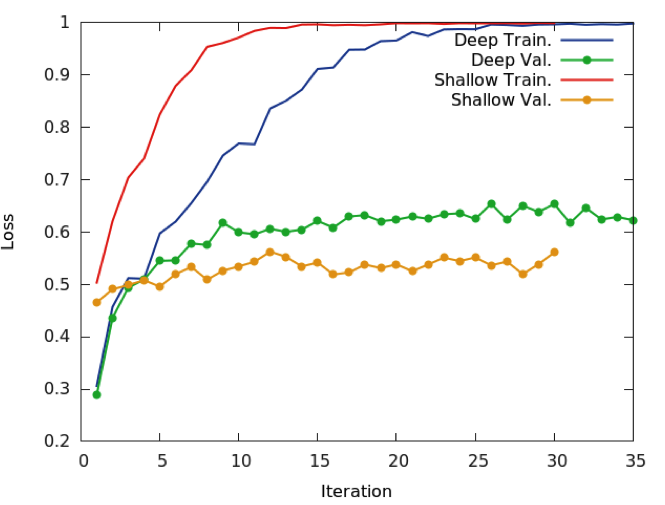


Fig 4: effectiveness: deep and shalllow model

utilized them as an in-put characteristic to the Fully connected layers in order to test the impact of using various features in our CNN model. We trained one deep network and one Shallow network using this concept. The acquired accuracy for the s deep and shallow

Fig 5: Confusion matrix : Shallow Model

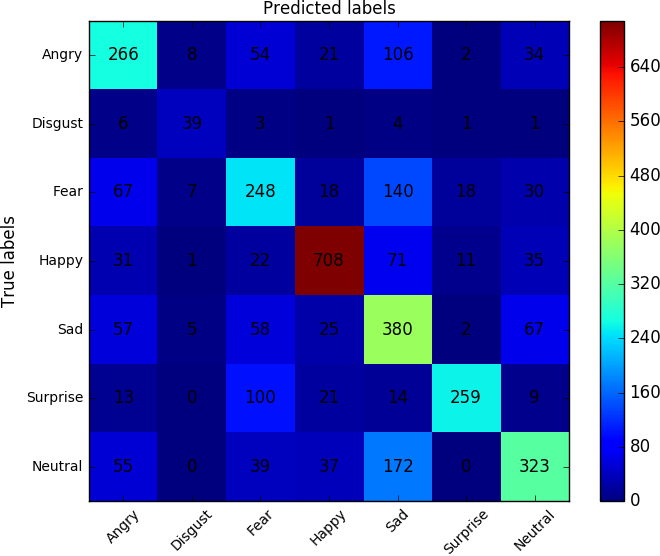


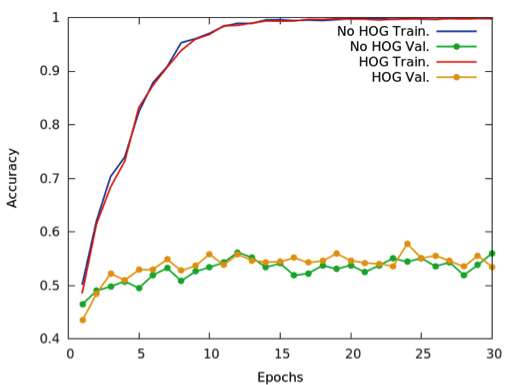
Fig 6: Confusion-matrix : Deep-model

The acquired accuracy for the deep and shallow models, respectively, is shown in Fig 8 and 7. These figures show that the model's accuracy is extremely close to the precision we attained with the model devoid of HOG components.This suggests that with only raw pixel data, CNN is capable of extracting sufficient data, including that obtained from HOG characteristics.

During the forward pass, we displayed the activation maps of several layers to track the features that each layer of our training set retrieves. This visualization is displayed in Figure 9. We may observe that the activation maps becoming sparser and more localized as the training goes on.

We also showed the weights of the first layer to show the qualification of the trained network. We offer quiet filters that don't produce noisy patterns., as shown in Figure 10. It shows that our network has received sufficient training, and the regularization strength is likewise adequate.

In order to uncover more complex patterns in our photos, we also applied the DeepDream [16, 17] technique to our top predictive model. Each expression's example and DeepDream output are shown in Figure 11 together.



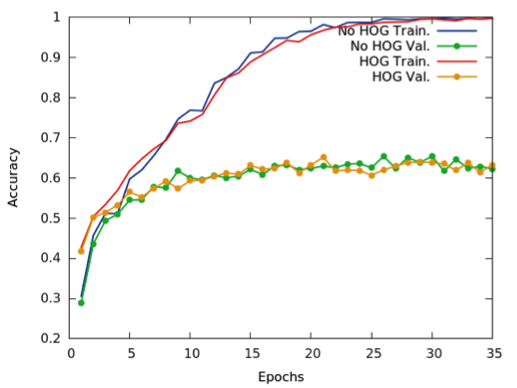
Fig 7: Accuracy of shallow model

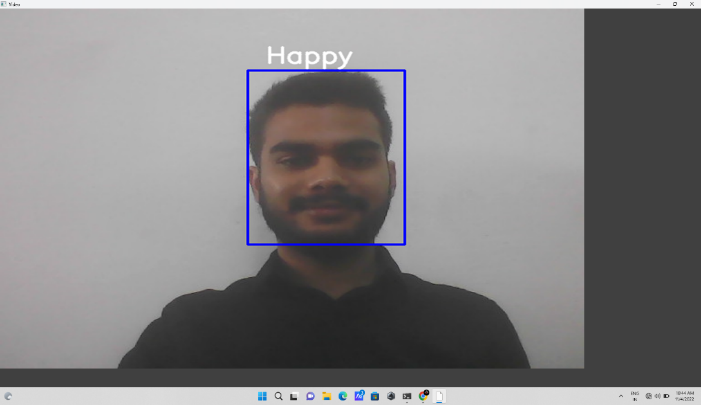
Fig 8: Accuracy of deep model

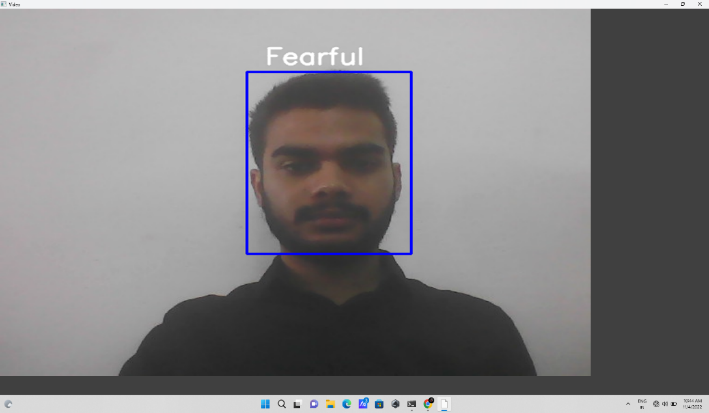
# Summary

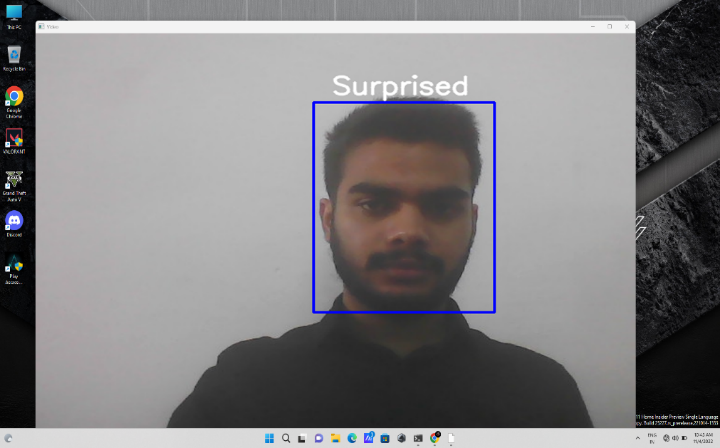
## Concluding

For the task of recognising facial expressions, we created a variety of CNNs, and we assessed their effectiveness using various post-processing and visualisation techniques. The outcomes showed that deep CNNs can learn facial characteristics and enhance face expression recognition. Additionally, the hybrid feature sets had little effect on the accuracy of the model, indicating that convolutional networks may naturally learn the main facial traits even when given simply raw pixel data.

Following are some images of the real time detection :







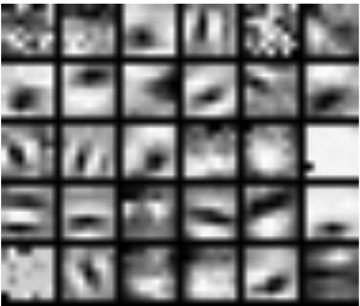
## Future Work

We used Torch's CNN packages to create all of the models for this project from scratch. In subsequent work, we hope to expand our model to include colour images.

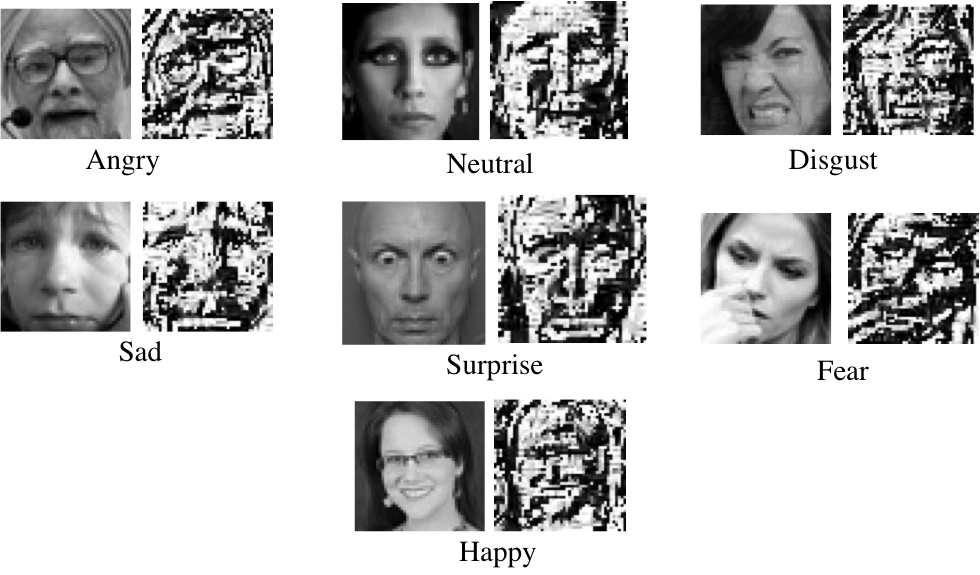


Fig 9: An illustration of our CNN's activation maps for several layers

This allow us to examine effectiveness of pretrained model for face emotion recognition, such as AlexNet or VGGNet. Implementing a face identification approach and then an emotion prediction mechanism would be another extension.









# References

1. [https://www.kaggle.com/c/challenges-in-](http://www.kaggle.com/c/challenges-in-) representation-learning-facial-expression- recognition-challenge/data

1. <https://github.com/torch>
2. Dalal, Navneet, and Bill Triggs (2005). Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on. Vol. 1
3. Bettadapura, Vinay (2012). Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722
4. Lonare, Ashish, and Shweta V. Jain (2013). A Survey on Facial Expression Analysis for Emotion Recognition. International Journal of Advanced Research in Computer and Communication Engineering 2.12
5. Nicu Sebe, Michael S. Lew, Ira Cohen, Yafei Sun, Theo Gevers, Thomas S. Huang (2007) Authentic Facial Expression Analysis. Image and Vision Computing 25.12: 1856-1863
6. Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), 2001.
7. M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In Proceedings of the IEEE ConFEDence on Automatic Facial and Gesture Recognition, 2006.
8. M. Pantic and J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. IEEE Transactions on Systems, Man and Cybernetics, 34(3), 2004
9. G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. Image and Vision Computing, 24(6), 2006.
10. M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J.R. Movellan. Automatic recognition of facial actions in spontaneous expressions. Journal of Multimedia, 2006.
11. P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978
12. Cohen, Ira, et al. ”Evaluation of expression recognition techniques.” Image and Video Retrieval. Springer Berlin Heidelberg, 2003. 184- 195
13. Padgett, C., Cottrell, G.: Representing face images for emotion classification. In: Conf. Advances in Neural Information Processing Systems. (1996) 894900.
14. <http://scikit-learn.org/stable/>

1. <http://deepdreamgenerator.com/>
2. <https://github.com/google/deepdream>
3. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ”Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems. 2012.
4. Simonyan, Karen, and Andrew Zisserman. ”Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556 (2014).
5. P. S Aleksic and A. K. Katsaggelos, “Automatic facial expression recognition using facial animation parameters and multistream hmms,” *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 1, pp. 3–11, 2006. View at: [Publisher Site](https://doi.org/10.1109/tifs.2005.863510) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Automatic%20facial%20expression%20recognition%20using%20facial%20animation%20parameters%20and%20multistream%20hmms&author=P.%20S%20Aleksic%20and%20A.%20K.%20Katsaggelos&publication_year=2006)
6. H. Chouhayebi, J. Riffi, M. A. Mahraz, Y. Ali, and T. Hamid, “Facial expression recognition using machine learning,” in *Proceedings of the 2021 Fifth International ConFEDence On Intelligent Computing in Data Sciences (ICDS)*, pp. 1–6, IEEE, Seoul, Korea, November 2021. View at: [Publisher Site](https://doi.org/10.1109/icds53782.2021.9626709) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Facial%20expression%20recognition%20using%20machine%20learning&author=H.%20Chouhayebi,%20J.%20Riffi,%20M.%20A.%20Mahraz,%20Y.%20Ali,%20and%20T.%20Hamid)