Affective Computing in Computer Vision: A Study on Facial Expression Recognition

Jordan J. Bird
Department of Computer Science
Nottingham Trent University
Nottingham, United Kingdom
jordan.bird@ntu.ac.uk

Azhar Aulia Saputra
Graduate School of Systems Design
Tokyo Metropolitan University
Hino, Tokyo, Japan
aa.saputra@tmu.ac.jp

Naoyuki Kubota
Graduate School of Systems Design
Tokyo Metropolitan University
Hino, Tokyo, Japan
kubota@tmu.ac.jp

Ahmad Lotfi
Department of Computer Science
Nottingham Trent University
Nottingham, United Kingdom
ahmad.lotfi@ntu.ac.uk

***Abstract*—The use of artificial intelligence has become increasingly popular in recent years, allowing technology once thought of as futuristic to become possible and utilised at the consumer level. Many technological barriers to human-computer interaction have been overcome, and there is now a focus on the sociological acceptance of such technology. Inferring human emotional states is a time-consuming process and can be automated with computer vision. In this study, we explore how computer vision and face recognition systems can be leveraged to automatically infer human emotional states from the face. Rather than the classical single-emotion classification method, our aim is to explore whether it is possible to perform regression techniques to observe valence and arousal. Following the topology tuning of 33 different neural networks, the results show that valence and arousal can be predicted by a branched Convolutional Neural Network model with a mean squared error of 0.066 and 0.107, respectively. In addition, we discuss methods of improving the model, as well as uses of the technology, which include the autonomous monitoring of affect during situations of technological acceptance.**

***Index Terms*—Affective Computing, Human-Computer Interaction, Computer Vision, Emotion Regression**

I. Introduction

In recent years, research and development of intelligent robotic platforms have seen a rapid increase, allowing technology once thought of as futuristic to become possible and used at the consumer level. Therefore, the barrier is not technological, but social, and the acceptance of such technologies is questionable. In this research, our aim is to explore emotional reactions to *artificially intelligent machines* through autonomous affective computing. There are many studies that explore the acceptance of new technology, often seen as radical by users [1]–[3]. Indeed, the acceptance of such technology could increase the efficiency and value of industrial domains such as customer service [4]–[6]. Given this, the need for affective reaction is paramount to understanding human behaviour within human-machine interaction, and its automation could enable studies en masse.

Classically, emotional recognition is often a concrete prediction of labels. For example, an image of a person smiling would be labelled as "Happy", and an image of a person crying would be labelled as "Sad". Though this is the case, human emotion is much more nuanced. For example, such a model given an input of a person crying with joy would likely encounter problems with labelling. Similarly, many emotions are similar; boredom and sleepiness cause the exhibition of similar facial expressions, as do surprised and scared states. This, then, suggests that Machine Learning (ML) approaches should learn to recognise minute physical differences, which would lead to distinctly different emotions. Human facial emotions are measured in psychology on a circumplex of two values, **valence** and **arousal**. Therefore, ML models should follow a similar approach to inference of the human face. In this study, we propose to predict the two values as a regression problem, rather than a single-class classification. This is achieved through computer vision, where each human face is extracted from an image before regression of the two values.

The scientific contributions of this work relate to the exploration of computer vision model topologies to predict emotional states. Following 33 tuning experiments, the final model was able to predict emotional states of valence and arousal with relatively high accuracy. The model was observed to achieve an average Mean Squared Error on validation data of 0.087. We open-source our model and make it available to the research community for future work[1].

The remainder of this article is as follows: Section II presents a review of the literature on related work in the field. Section III explains the methodology followed by a summary of the experiments carried out in this study, before the results

This work was supported by The Royal Society [Grant Number CR/212667] and the Japan Science and Technology Agency - Moonshot RnD [Grant Number JP- MJMS2034].

[1]The final model and Python code from this work is available at: https://github.com/jordan-bird/Valence-and-Arousal-Recognition-from-Human-Faces

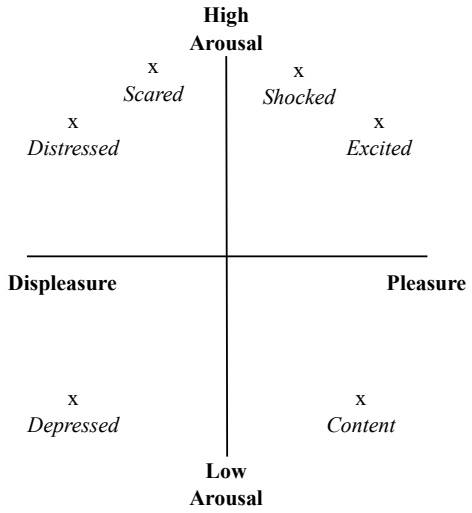

Fig. 1. An example of emotions on Russell's Circumplex Model of Affect [22].

are presented in Section IV. Finally, future work is discussed and this study is concluded in Section V.

II. Related Work

The related background work contributing to the understanding of human facial emotions, facial expression recognition, and the use of affective computing in human-machine interaction is presented in this section.

While Charles Darwin's work is most often related to biological evolutionary theories, his work [7] on *The Expression of the Emotions in Man and Animals* explored the psychological aspect of evolution. According to Darwin, particular instances of emotion share distinct and immutable states and are innate to all human beings. A large number of studies have suggested that universal facial expressions occur spontaneously during an emotional state [8], for example, studies on infants show that social smiles form around the same age regardless of whether the child is blind, visually impaired, or has 20:20 vision [9]. Given this, therein lies the argument that if some physiological emotional states are universal, then they could also be generalised. Facial Expression Recognition (FER) is the autonomous encoding and analysis of information expressed by the human face to derive affective information [10]. Facial emotion recognition is a subset of sentiment analysis that can be inferred from text [11], [12], human activity [13], [14], signal processing [13], [15], Computer Vision [16]–[18] and landmark classification [19]–[21].

Classically, emotion recognition was thought of as a classification-based task due to relatively scarce datasets [23]. Russell's Circumplex Model of Affect [22] instead suggests that emotion recognition is a more akin to a regression task; Russell argued that affective states arise from the neurophysiological systems of valence, or how pleasurable a state is, and arousal, a measure of how alert the state is. For example, the states of *surpised* and *scared* are similar in arousal, but differ in that one is of positive valence and the other negative. A diagram of the circumplex model can be found in Fig. 1. Mollahosseini, Hasani, and Mahoor presented the AffectNet dataset in 2017 [24], contributing a large-scale data set of 1 million images with values of valence and arousal, which allowed this task in machine learning. In the original paper, the authors found that AlexNet could achieve Root Mean Squared Error (RMSE) values of 0.37 and 0.41 for valence and arousal, respectively.

In 2019, Deng et al. [25] suggested a Conditional Generative Adversarial Network (cGAN) approach to facial expression recognition for Human-Robot Interaction. In this work, a cGAN was trained to change the emotion on facial expression images in order to alleviate intra-class variability. Although facial expressions have been argued to be universal and thus generalisable, such approaches can aid in situations of data scarcity (i.e., in the form of data augmentation). The approach was noted to be effective in decoupling variations of personal identity and pose from the emotional class. Melinte and Luige [26] proposed a framework which coupled region-based computer vision for face recognition and fine-tune transfer learning of Convolutional Neural Networks (CNN) for emotional recognition. It was noted that transfer learning from a pre-trained Residual CNN could recognise emotions with around 90.14% accuracy, and the model was deployed for use on Softbank's NAO robot.

In [27], the authors propose a computer vision-based approach to emotion recognition in the wild. Each of the seven basic emotional states was classified with a mean accuracy of around 61.29%; confusion matrices in this study show the confusion that models can have when classifying related emotional states. For example, more obvious emotions with high valence, such as anger and happiness, were relatively easier to classify compared to other emotions, such as fear, which was more often misclassified as surprise. As discovered during the literature review on the circumplex model of affect, emotions such as fear and surprise exist closely on the model and thus are exhibited in similar ways. Similarly, the studies in [28] discovered that there was crossover in predictions of anger and surprise in the Extended Cohn-Kanade dataset. Given this, it is therefore proposed to consider measures of valence and arousal rather than a single-label emotional state.

III. Method

In this section we discuss the methodology followed by the experiments in this work. Firstly, data collection and preprocessing of images will be presented prior to data selection and subsequent application of ML techniques.

The dataset for this study is collected from the AffectNet repository of facial images [24]. The dataset is comprised of 1 million images of human faces with perceived measures of valence and arousal on a scale of -1 to 1.

Fig. 2 shows four random examples of images from the dataset along with their perceived measurements of emotional valence and arousal. For this study, a random subset of 10,000 images are used to tune the model hyperparameters, before the

V: 0.78 A: -0.05 | V: -0.02 A: 0 | V: 0.55 A: 0.25 | V: 0.78 A: -0.14

Fig. 2. A sample of four random examples of valence and arousal measurements from the AffectNet dataset.

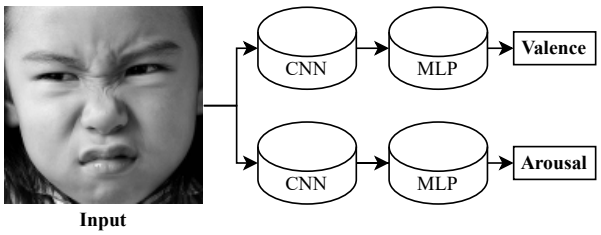

Fig. 3. Overview of the branched computer vision approach to affective regression.

final model is trained on 100,000 images selected at random. All images are resized as 128 px square images and colours are reduced to one (greyscale) channel. Thus, there are a total of 16,384 input parameters ($128 \times 128 \times 1$).

This study is supported by three main technologies. Firstly, the Convolutional Neural Network [29], [30], which is a subset of neural networks specialising in performing learnt operations on spatial data, in our case, implementing filters for visual data. CNNs are often three-dimensional in nature due to images often having three colour channels, though, in this work, images are one-channel (greyscale) and thus the CNN is 2-Dimensional. Secondly, the multilayer perceptron (MLP); similarly to the CNN, the goal of the neuronal layers are to extract underlying information from a given data, leading to an ideal transformation. The transformation in this case would be to take the outputs of the CNN features and predict a numerical value of affect.

Overall, the two CNNs and MLPs combine to form an overall system that (i) learns and extracts useful spatial features from images, (ii) processes these features to extract further higher-level information, and finally (iii) predicts two outputs based on the input face. The first being a measure of valence and the second being arousal, both on a scale of -1 to 1. This process of predicting a real number is known as *regression*.

During the initial benchmarking, it was discovered that a single computer vision model often produced poor regression results. This is possibly due to there being a lack of complementary features to contribute to both outputs. Therefore, instead, a branched CNN approach was selected, an example of which can be seen in Fig. 3. The input image is given to two individual neural networks which then produce outputs for both regression goals. The problem that the network is intended to solve is to reduce the mean difference between the predicted and real values. The general goal of the regression network is to reduce the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{1}$$

where, for n images, the MSE is the mean $\frac{1}{n}\sum_{i=1}^{n}$ of the errors squared $(Y_i - \hat{Y}_i)^2$ for real value Y_i and the predicted value $\hat{Y}_i$. Given that two values must be predicted for every data point, the loss function is therefore the mean of the two observed MSE values $\frac{MSE_v + MSE_a}{2}$, for the MSE of valence v and arousal a. Results are also presented in Root Mean Squared Error (RMSE) form since RMSE is measured with the same units as the output value, calculated as $RMSE = \sqrt{MSE}$. The expected outputs of the model are on a scale of -1 to 1 for both valence and arousal.

Each module of the system is tuned sequentially in grid search; initially, the CNN feature extractor is tuned with 1 to 4 layers, each comprised of 32, 64, or 128 neurons. Then, the secondary neural network is tuned from 1 to 3 layers each comprised of 32, 64, 128, 256, or 512 rectified linear units. Finally, Dropout [31] is introduced between each of the modules in a linear search of $\{0, 0.1, ..., 0.5\}$ to discover whether overfitting is occurring during the machine learning process. All machine learning models are trained using 80/20 data split for training and validation.

The weights of the neural network are trained with the ADAM optimiser [32] with a learning rate of 0.001, exponential decay for the first moment $\beta_1 = 0.9$, for the second moment $\beta_2 = 0.999$, and a constant for numerical stability $\epsilon = 1e-07$. All neurons are Rectified Linear Units (ReLu), $f(x) = \max(0, x)$, with the exception of the output neurons which are linear and have no activation function.

All benchmarking takes place on a subset of 10,000 images, and then the final model is trained on 100,000 images selected at random from the dataset. Due to this, statistical testing is required to observe the difference between the datasets, if any. To achieve this, we measure and compare the means, medians, and standard deviations of the observed emotional values. Following this, we then perform an unpaired t-test for both sets of valence and arousal values to discern whether there are statistical differences between the two. The models are trained via the TensorFlow 2.6.0 library on an Nvidia RTX 2080Ti GPU which has 11 GB of VRAM and 4,352 CUDA cores.

IV. Results and Discussion

The results of all experiments are presented in this section. Initially, the details of hyperparameter tuning experiments, which follow the process of tuning the CNN and MLP topologies, are presented before the implementation of dropout to prevent overfitting. Prior to training this topology on a larger set of data, statistical tests are then performed to observe any differences between the two datasets. Finally, the tuned topology is then trained on the full set of 100,000 images, and the overall results are presented.

TABLE I
MEAN LOSSES FOR THE FEATURE EXTRACTOR HYPERPARAMETER TUNING TOWARDS THE REGRESSION OF VALENCE AND AROUSAL FROM IMAGES OF FACIAL EMOTION.

Filters	**Layers**			
	1	***2***	***3***	***4***
32	0.154	0.14	0.124	0.114
64	0.154	0.137	0.122	**0.112**
128	0.153	0.139	0.128	0.113

TABLE II
MEAN LOSSES FOR THE TUNING OF DENSE LAYERS (FOLLOWING THE CNN) TOWARDS THE REGRESSION OF VALENCE AND AROUSAL FROM IMAGES OF FACIAL EMOTION.

Neurons	**Layers**		
	1	***2***	***3***
32	0.11	0.107	0.108
64	**0.106**	0.108	0.107
128	0.108	0.108	0.107
256	0.111	0.111	0.11
512	0.112	0.113	0.108

TABLE III
TUNING OF DROPOUT VALUES TO REDUCE OVERFITTING FOR THE REGRESSION OF VALENCE AND AROUSAL FROM IMAGES OF FACIAL EMOTION.

Dropout	**Validation Loss**
0	0.106
0.1	0.107
0.2	**0.105**
0.3	0.107
0.4	0.107
0.5	0.108

TABLE IV
STATISTICAL DIFFERENCES BETWEEN THE TWO DATASETS. THE 10K DATASET IS USED FOR TOPOLOGY TUNING, AND THE 100K DATASET IS USED FOR TRAINING THE FINAL MODEL.

Data	**Valence**			**Arousal**		
	Mean	**Median**	**Std.**	**Mean**	**Median**	**Std.**
10k	0.185	0.198	0.523	0.116	0.198	0.303
100k	0.186	0.203	0.517	0.118	0.068	0.302
Diff.	0.002	0.005	-0.006	0.002	-0.130	-0.001

A. Topology Tuning

For the first set of tests, the CNN-based feature extractor is trained without dense tertiary layers to discern the most effective topology within the selected bounds. Table I shows the results and it can be observed that four layers of 64 filters produced the strongest model, resulting in the lowest mean loss of 0.112. This network is then used as the basis for the subsequent dense layers.

Table II shows the results for dense networks, which are attached to the CNN output. Within these results, it can be observed that the best network was comprised of the CNN feeding into a single layer of 64 rectified linear units. This led to a lower mean loss of 0.106.

In the final topology tuning experiments, Table III shows the tuning of dropout values from 0.1 to 0.5. A slightly lower mean loss of 0.105 was achieved given a 20% dropout rate between each layer, suggesting that some overfitting was taking place during learning.

The general topology of the model defined by the described experiments was a CNN feature extractor comprised of 4 layers of 64 filters, which was then attached to a dense layer of 64 rectified linear units, each layer had a dropout rate of 20%.

B. Statistical Tests between Datasets

Given that topology tuning was trained on a set of 10,000 images due to the limited availability of computational resources, and the final model was then trained on a larger set of 10,000 images, statistical differences between the two sets must be observed. Firstly, Table IV shows the statistics of each dataset. Here we can see that the sets of model outputs are statistically similar, with the mean valence changing by 0.001 between the datasets. Similarly, the difference in the means of the arousal states had a difference of 0.002. For valence, the standard deviation of the 10k dataset was 0.523 and was 0.517 for the 100k dataset, resulting in a difference of 0.006. For arousal, the difference was smaller, with a change of 0.001 between the two sets. Table V shows an unpaired t-test between the expected model outputs for the two datasets. Given p of 0.369 for valence and 0.29 for arousal, it can be argued that there is no statistical significance between the datasets. Thus, the tuned topology is then trained on the full set of images.

C. Training the Final Model

Table VI shows the results for the final model. The chosen topology is that which was derived from the tuning experiments when considering 10,000 images, and these results show the metrics for 100,000 images. The RMSE for valence was observed to be around 0.257, and 0.327 for arousal[2].

V. FUTURE WORK AND CONCLUSION

In this study, we have explored a method to predict human emotional states from the face with computer vision. With relatively few computational resources training a small model, it was observed that it is possible to predict both the valence and arousal states of the emotion. In future, given more resources, the model could be trained on the full set of 1 million images. Dimensions were reduced to greyscale, but skin colouration may be useful for the recognition of affect as it is to measure heart rate in video [33]; therefore, future work could concern the regression of RGB images. Following further tuning, the model could be used for case studies involving technological acceptance studies. For example, when interacting with humanoid robots, the deployment of this

[2]The final model is open source and can be downloaded from GitHub: https://github.com/jordan-bird/Valence-and-Arousal-Recognition-from-Human-Faces

TABLE V
RESULTS OF THE UNPAIRED T-TEST BETWEEN THE SET OF 10K IMAGES USED FOR TOPOLOGY TUNING AND THE SET OF 100K IMAGES FOR TRAINING THE FINAL MODEL.

	Valence	**Arousal**
p-value	0.369	0.290

TABLE VI
FINAL RESULTS WHEN THE FINAL MODEL IS TRAINED ON A LARGER DATASET OF 100,000 IMAGES.

	MSE	**RMSE**	**Loss**
Valence	0.066	0.257	0.108
Arousal	0.107	0.327	0.0672
Mean	0.0865	0.292	0.0876

model would allow both the robot and observers to measure the change in emotion during interactive activities. Similarly, the model has relevant applications in consumer studies, such as when a potential customer interacts with a product or consumes media. In our future work, we plan on applying the model arising from this study in Human-Robot Interaction to study human emotional reactions to technology in real-time.

To finally conclude, the results arising from this work show that it is possible to predict human emotional states on the numerical scale of Russell's circumplex using computer vision. Following the topology tuning of 33 individual neural networks, the final results shows that valence and arousal could be predicted from the human face with mean squared errors of 0.066 and 0.107, respectively.

REFERENCES

[1] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler, "An experimental study on emotional reactions towards a robot," *International Journal of Social Robotics*, vol. 5, no. 1, pp. 17–34, 2013.

[2] R. Bevilacqua, M. Di Rosa, G. R. Riccardi, G. Pelliccioni, F. Lattanzio, E. Felici, A. Margaritini, G. Amabili, and E. Maranesi, "Design and development of a scale for evaluating the acceptance of social robotics for older people: The robot era inventory," *Frontiers in neurorobotics*, vol. 16, 2022.

[3] U. A. Saari, A. Tossavainen, K. Kaipainen, and S. J. Mäkinen, "Exploring factors influencing the acceptance of social robots among early adopters and mass market representatives," *Robotics and Autonomous Systems*, vol. 151, p. 104033, 2022.

[4] J. Wirtz, P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch, and A. Martins, "Brave new world: service robots in the frontline," *Journal of Service Management*, 2018.

[5] A. Tuomi, I. P. Tussyadiah, and J. Stienmetz, "Applications and implications of service robots in hospitality," *Cornell Hospitality Quarterly*, vol. 62, no. 2, pp. 232–247, 2021.

[6] A. Tuomi, I. P. Tussyadiah, and P. Hanna, "Spicing up hospitality service encounters: the case of pepper™," *International Journal of Contemporary Hospitality Management*, 2021.

[7] C. Darwin, "The expression of the emotions in man and animals (1872)," *The Portable Darwin*, pp. 364–393, 1993.

[8] D. Matsumoto, D. Keltner, M. N. Shiota, M. O'Sullivan, and M. Frank, "Facial expressions of emotion." 2008.

[9] S. J. Rogers and C. B. Puchalski, "Social smiles of visually impaired infants," *Journal of Visual Impairment & Blindness*, vol. 80, no. 7, pp. 863–865, 1986.

[10] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[11] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Emotion measurement*. Elsevier, 2016, pp. 201–237.

[12] J. J. Bird, A. Ekárt, C. D. Buckingham, and D. R. Faria, "High resolution sentiment analysis by ensemble classification," in *Intelligent Computing-Proceedings of the Computing Conference*. Springer, 2019, pp. 593–606.

[13] G. Cai, X. He, and J. Pan, "Visual sentiment analysis with local object regions attention," in *International Conference of Pioneering Computer Scientists, Engineers and Educators*. Springer, 2019, pp. 479–489.

[14] A. M. Sadiq, H. Ahn, and Y. B. Choi, "Human sentiment and activity recognition in disaster situations using social media images based on deep learning," *Sensors*, vol. 20, no. 24, p. 7115, 2020.

[15] A. Samareh, Y. Jin, Z. Wang, X. Chang, and S. Huang, "Detect depression from communication: how computer vision, signal processing, and sentiment analysis join forces," *IISE Transactions on Healthcare Systems Engineering*, vol. 8, no. 3, pp. 196–208, 2018.

[16] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with cnn ensemble," in *2016 international conference on cyberworlds (CW)*. IEEE, 2016, pp. 163–166.

[17] M. Shin, M. Kim, and D.-S. Kwon, "Baseline cnn structure analysis for facial expression recognition," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2016, pp. 724–729.

[18] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, and L. Xun, "Facial expression recognition with faster r-cnn," *Procedia Computer Science*, vol. 107, pp. 135–140, 2017.

[19] D. R. Faria, M. Vieira, F. C. Faria, and C. Premebida, "Affective facial expressions recognition for human-robot interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 805–810.

[20] M. Munasinghe, "Facial expression recognition using facial landmarks and random forest classifier," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. IEEE, 2018, pp. 423–427.

[21] Y. Qiu and Y. Wan, "Facial expression recognition based on landmarks," in *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, vol. 1. IEEE, 2019, pp. 1356–1360.

[22] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[23] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.

[24] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[25] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "cgan based facial expression recognition for human-robot interaction," *IEEE Access*, vol. 7, pp. 9848–9859, 2019.

[26] D. O. Melinte and L. Vladareanu, "Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified adam optimizer," *Sensors*, vol. 20, no. 8, p. 2393, 2020.

[27] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435–442.

[28] S. Xie and H. Hu, "Facial expression recognition with frr-cnn," *Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.

[29] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.

[30] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *2017 international conference on communication and signal processing (ICCSP)*. IEEE, 2017, pp. 0588–0592.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

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

[33] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.