DeepFER: Facial Expression Recognition with Deep Neural Network and Attentional Convolutional Network

Mubashir Mohsin [1] (id: 20-42884-1) [a], Shahriar Hossain Rafi [2] (id: 20-42528-1)[a], Jishan Ferdows Navil [3] (id:20-42544-1)[a]

[a]*Department of Computer Sciences, American International University-Bangladesh*

ABSTRACT

This paper delves comprehensively into the domain of facial expression recognition, a subject of burgeoning interest in recent decades. The inherent difficulty stems from the substantial intraclass variety seen in the field of facial expressions. In traditional approaches, classifiers trained on extensive image or video datasets are frequently combined with manually produced features like SIFT, HOG, and LBP. While these techniques have shown impressive effectiveness when used on controlled picture datasets, their performance deteriorates when applied to more intricate data sets with more image variance and uncompleted facial information. The ascendancy of deep learning models has ushered in a paradigm shift by ushering in end-to-end frameworks for facial expression recognition. Despite their discernible strides in performance, a substantial scope for refinement remains unexplored. This study proposes a pioneering deep learning framework hinged on an attentional convolutional network. This innovative architecture attains the ability to selectively focus on pivotal facial components, eliciting a pronounced enhancement over preceding models across diverse datasets, encompassing FER-2013, CK+, FERG, and JAFFE. A visualization technique is also developed that, based on the predictions of the classifier, identifies crucial facial regions that are influential in the identification of various emotional states. Experimental data underwent an empirical examination, which affirmed that various emotions display distinct reactions to various facial regions.

Keywords: ***Deep learning, Computer Vision, Facial Expression Recognition, Spatial Transformation, Attention Mechanism, Neural Network, CNN***

INTRODUCTION

Facial expressions are a significant point of human non-verbal communication, serving as a conduit for passing on assumptions, excitedly, and social signals. The capacity to precisely recognize facial expressions holds essential suggestions for a wide cluster of spaces, checking human-computer interaction, enthusiastic computing, brain ask approximately, and without a question law prerequisite [3]. Adjusted facial expression assertion frameworks energize the extraction of enthusiastic states from people, empowering a more noteworthy understanding of their energized reactions and supporting in making compassionate AI frameworks. Facial expressions are enthusiastic, advancing shapes that show up over time. Capturing the worldly stream of facial expressions is basic for a comprehensive

understanding of assumptions. In extension, not all facial zones contribute furthermore to communicating a particular feeling. For occasion, the mouth locale can be more educator for recognizing elation, whereas the eyebrow and haven locales might play an essential parcel in recognizing stun. Convolutional Neural Systems (CNNs) have laid out unprecedented execution in picture classification assignments, persuading their application to facial expression assertion [4].

In this paper, we appear a novel approach to facial expression confirmation, the Attentional Convolutional Organize (ACN), which opens up the capabilities of CNNs by joining thought components to center on the striking divide of the stand up to. This Paper Talks to:

[I] In advancement, we utilize the visualization strategy proposed in [5] to highlight the go up against image's most striking areas.

[II] the parts of the picture which have the foremost grounded influence on the classifier's result.

Interior the taking after parts, we detail the arrange, arranging strategy, and test comes around of the proposed ACN. We appear its execution on benchmark datasets and compare it to existing CNN-based models. Also, we dive into the interpretability of ACN's thought maps and see at their recommendations.

RELATED WORKS:

Prior research in facial expression recognition has largely focused on feature extraction, representation learning, and model design. In the earlier works on facial recognition some features were extracted and then an old Machine learning classifier was used to detect emotions. Histogram of oriented gradients (HOG) [8], [6], local binary patterns (LBP) [7], These approaches were working fine until the advance datasets came who have their own limitations like intra class variation. But with the introduction of deep learning many Facial Expression Recognition models have been introduced. Xiangyun Zhao [9] explained novel peak-piloted deep network for facial expression recognition. The main novelty is the embedding of the expression evolution from non-peak to peak into the network parameters. Yoshihiro Shima [10] proposed a combination of a deep neural network and a support vector machine. A neural network pre-trained with a large-scale object image database was used as a feature extractor for facial images. Mariana-Iuliana Georgescu [11] presented state-of-the-art approach for facial expression recognition, which is based on combining deep and handcrafted features and on applying local learning in the training phase.

The work mentioned above considerably improved facial expression recognition. Those works do, however, have some drawbacks. In this research, we present a model based ACNN to address such shortcomings by focusing on important picture regions.

PROPOSED METHODOLOGY

The project proposes an end-to-end deep-learning framework based on the facial features of human facial expressions to classify the underlying emotions in the face images. In computer vision and image processing with deep learning, it often comes to the point when a neural network is built,

adding more hidden layers and neurons improves the performance. Even adding normalization, regularizations and augmentation improves the results. But for facial expression recognition it is not always applicable because of the similarity in the salience area, or a salience features. The facial features of several emotions are carried out by similar facial points which leads to difficulty even for human eyes to differentiate between different emotional states. Due to the small number of classes in the FER, our model shows promising results using less than 10 layers. The model has been trained from scratch and able to achieve promising results in different FER-based datasets.

An emotional face image does not need all the parts of the face to be recognized for classification of an emotion. Based on the facial feature extraction of emotions (Elham et. al), the local features on the face such as nose, eyes, mouth are extracted in order to utilize the shapes and geometric analysis of different emotions. Therefore, our model added an attention mechanism, through the learning of spatial transformer network (Abdolrashidi and Minaee). This transformer network was used in our model to focus on important face regions rather than scanning over the whole image of the face.

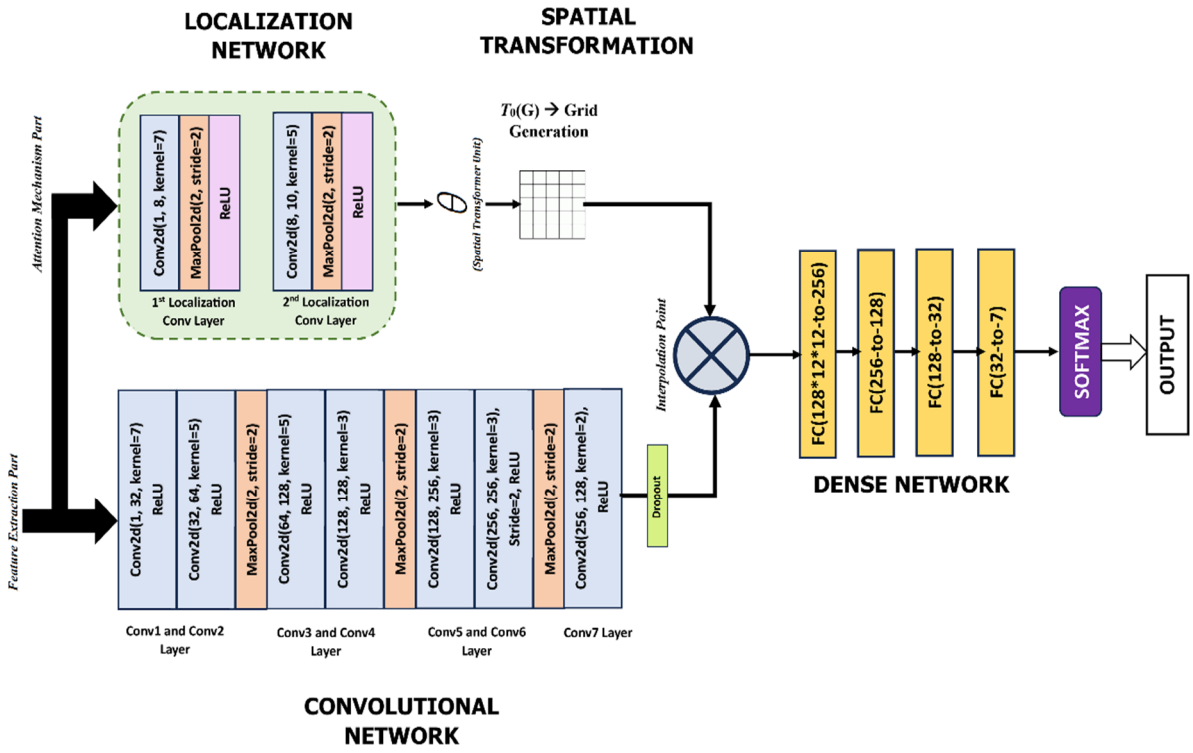


Fig 1: The proposed model architecture (DeepFER)

Figure 1 illustrates our proposed model (DeepFER) architecture with detailed information with layers. The whole architecture has been divided into two parts – Feature Extraction Part and Attention Mechanism Part. The feature extraction part consists of seven convolutional layers, each two except the last one followed by max-pooling layer and rectified linear unit (ReLU) activation function. They are then followed by a dropout layer which leads to the interpolation point. In the attention mechanism part, the regression of transformation parameters takes place, then the input is transformed to the sampling grid $\boldsymbol{\tau(\theta)}$ to produce the warped data. By determining a sample over the observed locale, the spatial transformer module effectively tries to focus on the image's most prominent area. To warp the input to the output, there are a variety of different transformations that may be utilized, but in this case, we chose an affine transformation, which is often used in many applications. For further details on

spatial transformer networks, see (Jederberg et al.). A loss function is then optimized using Adam optimizer to train this model. In this study, the classification loss (cross-entropy) and regularization term, which is the ℓ_2 normalization of the weights in the last four fully connected layers, are simply added to form the loss function.

$$\mathcal{L}_{overall} = \mathcal{L}_{classifier} + \lambda \|\omega(f_c)\|_2^2 \quad \ldots (1)$$

The validation set serves as the basis for tuning the regularization weight. We train our model entirely from scratch by including both dropout and ℓ_2 regularization. As a result, we were able to train our models entirely from scratch on small datasets like JAFFE and CK+. It is also important to note that for each dataset utilized in this study, a unique model was trained. We made use of the model from the paper we worked on, which had fewer convolutional layers and less intricate parameters. The test results for all datasets, however, improved as a result of our development of the proposed model.

EXPERIMENTAL RESULTS

We present a thorough experimental analysis of our model on various facial expression databases in this part. We first give a quick description of the databases we employed in this work, after which we show how well our models performed on four different databases and contrast the findings with some recent, exhilarating research. Then, using a visualization method, we present the prominent regions identified by our trained model. Several well-known facial expression recognition datasets, including FER2013 (Carrier, P. L. et al.), the expanded Cohn Kanade (CK+) (Lucey, Patrick, et al.), and Japanese Female Facial Expression (JAFFE) (Lyons, et al.) are used in this work to offer the experimental examination of the proposed approach. We'll give a quick rundown of these databases before getting into the findings.

JAFFE: This dataset includes 213 pictures of the seven facial expressions that ten Japanese female models were asked to make. Six emotion adjectives were used to score each image by 60 Japanese participants (Lyons, et al.) Figure 2 displays four examples of the dataset's photos.

Fig 2: Four samples from JAFFE dataset

For each image, we applied seven different data augmentation strategies for our model to improve performance, and we only clipped the face portion of the image. As a result, 1335 total images were produced, and we divided the data into three parts based on those images: 935 train samples, 140 validation samples, and 260 test samples. After training, we achieved test accuracy and validation accuracy of 89.38% and 91.15%, respectively. Saliency mapping then increased these results to validation accuracy and test accuracy of 92.13% and 93.89%, respectively.

CK+: A publicly available dataset for action unit and emotion recognition, the extended Cohn-Kanade (also known as CK+) facial expression database (Lucey, Patrick, et al.) is used. It comprises both posed and unposed (spontaneous) expressions. CK+ contains 123 different subjects and a total of 981 sequences. Figure 3 displays six representative pictures from this collection.

Fig 3: Four samples from CK+ dataset

This dataset has more samples and a more exact section of the facial image, which is ideal for training the model. As a result, we obtained 99.47% test accuracy and an average validation accuracy of 97.79%. We did not test any additional augmentation or saliency mapping techniques because this was producing nearly flawless results on the main dataset. The CK+ dataset will, however, yield the best results for this model.

FER2013: First introduced in the ICML 2013 Challenges in Representation Learning (Carrier, P. L. et al.), the Facial Expression Recognition 2013 (FER2013) database. This dataset contains 35,887 images of 48x48 resolution, most of which were taken in wild settings. 28,709 photos made up the initial training set, while 3,589 images each made up the validation and test sets. Faces are automatically captured in this database because the Google image search API was used to generate them. The six cardinal expressions, as well as neutral, are used to label faces. FER has more intra-class variation in the photos than the other datasets, including half faces, low-contrast images, eyeglasses, and face blockage (most often with a hand). In the training folder, there could potentially be hundreds of photographs that have incorrect labels. Due to this, it is extremely difficult to recognize facial points, transform, or tune into the database. Fourteen images from the FER dataset are shown in Figure 4.

Fig 4: Fourteen samples from FER2013 dataset

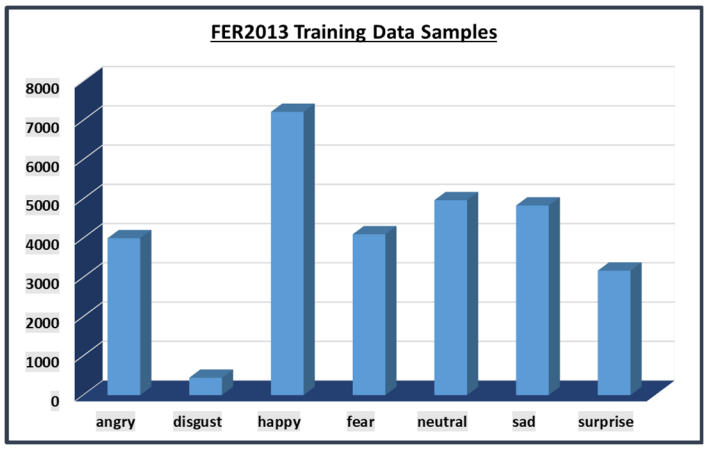


Fig 5: FER2013 dataset training sample distribution

In this dataset, the data sample is not the same for each class as shown in figure 5. The disgust class has only 436 photos, which is drastically imbalanced for an image classification model where each class contains about 3000 images. So, using this dataset, we performed two separate evaluations. First, we randomly took 430 images (400 for train and 30 for validation) from each class of the train folder

and 100 images from each class of the test data. Total: 2800 samples for training, 210 samples for validation, and 700 samples for testing. Total: 2800 training samples, 210 validation samples, and 700 testing samples. Our model achieved a test accuracy of 74.23% and a validation accuracy of 69.76% using this sample dataset. The model was then evaluated on the complete dataset, with test and validation accuracy of 72.33% and 68.06%, respectively. However, as nearly half of the images had occlusion, face blocking, or were half-faced, we were unable to employ any saliency mapping procedures on this dataset.

FERG: A database of stylized characters with annotated facial emotions is known as FERG. Six stylized characters are represented by 55,767 annotated images of faces in the database. MAYA was used to model the characters. Seven categories of expressions are used to organize the photos for each character [2]. Figure 6 displays six representative pictures from this database. The major reason we tested our algorithm on this database was to evaluate how it handled cartoon characters.

Fig 6: Six samples from FERG dataset

This dataset offers a large number of balanced samples, which is helpful for model training. However, as they are cartoons, all of the characters have the same facial features, regardless of the character's emotion. Considering this, there is a strong likelihood that the entire dataset will show very impressive accuracy. Due to this, we used two separate evaluation techniques in this instance as well. Out of the six characters, we first chose just four for the training samples: Aia, Bonnie, Jules, and Malcolm. We chose the final two characters, Mery and Ray, for the test samples. When we ran our model, we obtained exemplary outcomes of 97.21% validation accuracy and 98.62% test accuracy. Following that, we took into account all of the characteristics for the second evaluation and divided the data into a trio, totaling 36,500 samples for training, 12,620 samples for testing, and 6,647 samples for validation. This dataset outperformed the other dataset by having a flawless test accuracy of 100% (tested three times with three distinct epochs with the same accuracy) and an average validation accuracy of 99.97%. This wasn't what would have been expected, but it makes sense given the variation less face and cartoon-face photos produced by MAYA software. Therefore, it is much simpler for an image classification model to determine facial orientation because each and every character's face in the image has the same forms, shades, contrast, brightness, sharpness, and hardness.

The research paper we selected (Abdolrashidi and Minaee) was a state-of-the-art framework for FER datasets and an emotion recognition model. However, we initially applied the author's model to our dataset and obtained results that were almost identical to those reported in the research, with the exception of the JAFFE dataset. On the initial 213 photos in the JAFFE dataset, the author asserts that their model was able to improve test accuracy by 92.8%. With several assessments using a variety of dataset selection techniques on the JAFFE dataset, we were unable to get any better than 85.88% using that model. However, we believe there could have been an oversight in the selection of the dataset or in the execution of the programs from our side. So perhaps that is why we were not able to achieve the publicized accuracy.

The performance of our model in comparison to other well-known FER-based models that have been supported up to this point is shown in the following tables and the confusion matrix of each dataset evaluation.

Classification Accuracy Comparison:

Method	Accuracy Rate
Bag of Words (Ionescu, Radu et al.)	67.40%
VGG+SVM (Georgescu, Iuliana et al.)	66.31%
Deep-Emotion (Abdolrashidi and Minaee)	70.02%
Proposed Method	**74.85%**

Accuracies on FER 2013 dataset

Method	Accuracy Rate
ST+RNN (Zhang, T. et al.)	97.20 %
PPDN (Zhao, Xiangyun et al.)	97.30%
IACNN (Meng, Zibo, Ping, L. et al.)	95.37%
IB-CNN (Han et al.)	95.10%
Deep-Emotion (Abdolrashidi and Minaee)	98.00%
Proposed Method	**99.47%**

Accuracies on CK+ dataset

Method	Accuracy Rate
CNN+SVM (Shima, Yoshiro et al.)	**95.31%**
Salient Facial Patch (Happy, S. L et al.)	91.80%
Fisherface (Abidin, Zaenal et al.)	89.20%
Deep-Emotion (Abdolrashidi and Minaee)	92.80%
Proposed Method	93.89%

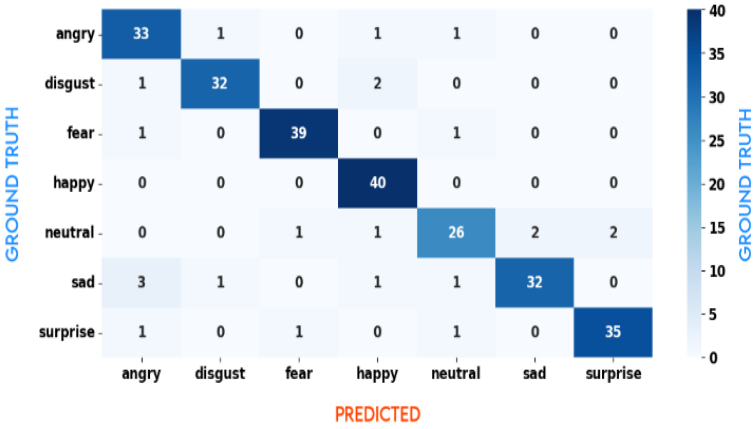
Accuracies on JAFFE dataset

Method	Accuracy Rate
DeepExpr (Aneja, Deepali et al.)	89.02%
Ensemble Multi-feature (Hang, Zhao et al.)	97.00%
Adversarial NN (Feutry, Clment, Pablo et al.)	98.20%
Deep-Emotion (Abdolrashidi and Minaee)	99.30%
Proposed Method	**100.00%**

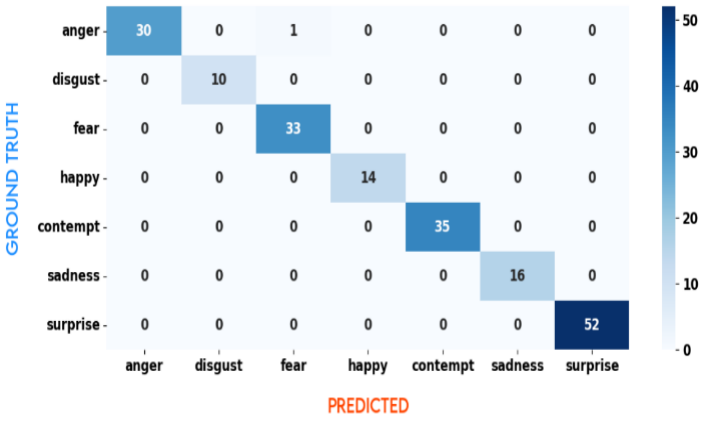
Accuracies on FERG dataset

Confusion Matrix:

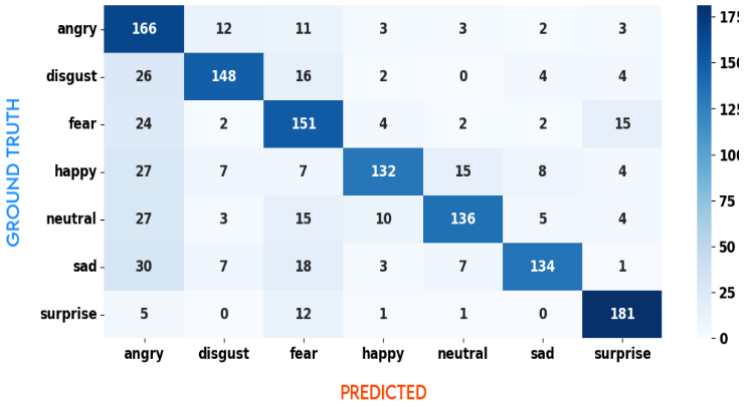
JAFFE Dataset



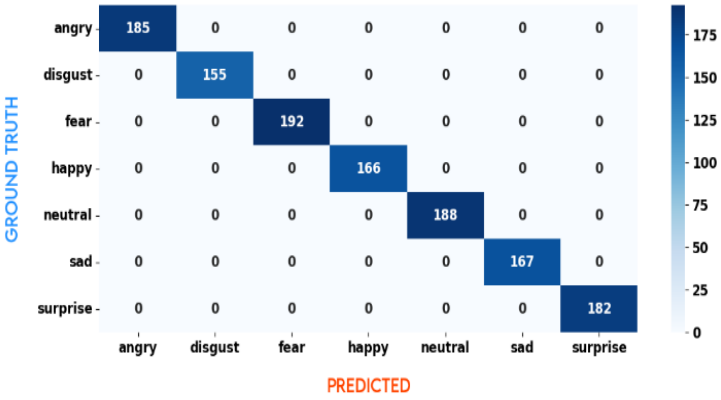
CK+ Dataset



FER2013 Dataset



FERG Dataset



DISCUSSION

The proposed methodology for facial expression identification offers a thorough strategy for addressing the complexities of facial feature-based emotion classification. The difficulties brought on by the commonality in salient facial expressions denoting different emotional states are addressed by DeepFER, an end-to-end deep-learning system. Contrary to normal computer vision tasks, the shared face traits of emotions mean that performance isn't always improved by increasing model complexity. In order to accommodate the peculiarities of emotion recognition, this project carefully chooses a model with fewer than 10 layers and prioritizes successful feature extraction over sheer depth. The incorporation of an attention mechanism through a spatial transformer network is the key component of the suggested approach. Critical face regions are strategically identified by the process, allowing for concentrated analysis and feature extraction. This technique dramatically enhances the model's capacity to recognize minute emotional cues in particular facial regions.

The uneven class distribution in the widely used dataset FER2013 presented problems. The model achieved promising validation and test accuracies by carefully partitioning and enhancing the dataset.

However, the usefulness of saliency mapping techniques was constrained by the existence of occlusions and various facial angles. When compared to a state-of-the-art model, the proposed method produced competitive results across multiple datasets. The two models' performance differences on the JAFFE dataset raise issues regarding the selection and use of the dataset. The study focuses on applying sophisticated methods and attention processes to recognize emotions in a complex environment. By fusing intelligent model architecture with attention processes, it increases the accuracy of emotion categorization. Future developments will include studying performance disparities and enhancing saliency mapping techniques for datasets containing occlusions. This study establishes a strong basis for improved emotion identification models.

CONCLUSIONS

This study has provided a thorough and efficient method for using deep learning to recognize facial expressions. By incorporating an attention mechanism through spatial transformer networks, the model is better able to identify key facial regions, increasing the accuracy of emotion classification. The success of the strategy is supported by experimental findings from several facial expression databases. Notably, the model displayed remarkable performance on datasets including JAFFE, CK+, FER2013, and FERG. The study demonstrated the approach's shortcomings, especially in datasets with occlusions and a range of facial angles, while still reaching outstanding accuracy. Through the careful use of attention mechanisms and an emphasis on effective feature extraction over pure model depth, this research advances the field of emotion identification models. Future study could improve the methodology's applicability and accuracy in a wider range of settings by exploring more advanced saliency mapping approaches and delving into dataset-specific performance differences.

REFERENCES

[1] S. Minaee, “Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network,” arXiv.org, Feb. 04, 2019. [Online]. Available: https://arxiv.org/abs/1902.01019

[2] V.Upadhyay, D. Kotak “A Review on Different Facial Feature Extraction Methods for Face Emotions Recognition System,” IEEE Conference Publication | IEEE Xplore, Jan. 01, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9171172

[3] Han, Shi Zhong, Zibo Meng, Ahmed-Shehab Khan, and Yan Tong. “Incremental boosting convolutional neural network for facial action unit recognition.” In Advances in Neural Information Processing Systems, pp. 109-117. 2016

[4] Meng, Zibo, Ping Liu, Jie Cai, Shi Zhong Han, and Yan Tong. “Island Loss for Learning Discriminative Features in Facial Expression Recognition.” In IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2017

[5] Happy, S. L., and Aurobindo Routray. “Automatic facial expression recognition using features of salient facial patches.” IEEE transactions on Affective Computing 6.1: 1-12, 2015

[6] Hough, Paul VC.” Method and means for recognizing complex patterns.” U.S. Patent 3,069,654, issued December 18, 1962

[7] Shan, Caifeng, Shaogang Gong, and Peter W. McOwan.” Facial expression recognition based on local binary patterns: A comprehensive study.” Image and vision Computing 27.6: 803-816, 2009

[8] Chen, Junkai, Zenghai Chen, Zheru Chi, and Hong Fu.” Facial expression recognition based on facial components detection and hog features.” In International workshops on electrical and computer engineering subfields, pp. 884-888, 2014

[9] Zhao, Xiangyun, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan.” Peak-piloted deep network for facial expression recognition.” In European conference on computer vision, pp. 425-442. Springer Cham,2016.

[**10**] Shima, Yoshihiro, and Yuki Omori.” Image Augmentation for Classifying Facial Expression Images by Using Deep Neural Network Pre-trained with Object Image Database.” Proceedings of the 3rd International Conference on Robotics, Control and Automation. ACM, 2018

[**11**] Georgescu, Mariana-Iuliana, Radu Tudor Ionescu, and Marius Popescu.” Local Learning with Deep and Handcrafted Features for Facial Expression Recognition.” arXiv preprint arXiv:1804.10892, 2018.

[**12**] Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor.” Emotion recognition in human-computer interaction.” IEEE Signal processing magazine 18, no. 1: 32-80, 2001.

[**13**] Aneja, Deepali, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. “Modeling stylized character expressions via deep learning.” In Asian Conference on Computer Vision, pp.136-153. Springer, Cham, 2016.

[**14**] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. “Spatial transformer networks.” Advances in neural information processing systems, 2015.

[**15**] Lucey, Patrick, et al. “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression.” Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference on. IEEE, 2010

[**16**] Carrier, P. L., Courville, A., Goodfellow, I. J., Mirza, M., & Bengio, Y. FER-2013 face database. Universit de Montreal, 2013

[**17**] Lyons, Michael J., Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. “The Japanese female facial expression (JAFFE) database.” third international conference on automatic face and gesture recognition, pp. 14-16, 1998.

[**18**] Zhang, T., Zheng, W., Cui, Z., Zong, Y., & Li, Y. Spatialtemporal recurrent neural network for emotion recognition. IEEE transactions on cybernetics, (99), 1-9, 2018.

[**19**] Abidin, Zaenal, and Agus Harjoko. "A neural network based facial expression recognition using fisherface." International Journal of Computer Applications 59.3, 2012

[**20**] Zhao, Hang, Qing Liu, and Yun Yang. "Transfer learning with ensemble of multiple feature representations." 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA), IEEE, 2018.

[**21**] Ionescu, Radu Tudor, Marius Popescu, and Cristian Grozea. "Local learning to improve bag of visual words model for facial expression recognition." Workshop on challenges in representatio0n learning, ICML. 2013.

[22] Elham Bagherian, Rahmita.Wirza.Rahmat and Nur Izura Udzir"Extract of Facial Feature Point" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.1, January 2009.

CONTRIBUTION

	Mubashir Mohsin *20-42884-1*	**Shahriar Hossain Rafi** *20-42528-1*	**Jishan Ferdows Navil** *20-42544-1*	**Contribution (%)**
Conceptualization	35%	35%	30%	100 %
Data curation	0%	40%	60%	100 %
Formal analysis	20%	30%	50%	100 %
Investigation	15%	70%	15%	100 %
Methodology	90%	10%	0%	100 %
Implementation	80%	10%	10%	100 %
Validation	60%	30%	10%	100 %
Theoretical derivations	40%	40%	20%	100 %
Preparation of figures	30%	0%	70%	100 %
Writing – original draft	5%	70%	25%	100 %
Writing – review & editing	40%	20%	40%	100 %