Emotion Recognition by using optimised deep features

Irfan Haider, Guee-Sang Lee*, Hyung-Jeong Yang, and Soo-Hyung Kim

Department of Artificial Intelligence Convergence
Chonnam National University, Gwangju 61186, South Korea
irfan_haider99@hotmail.com, gslee@jnu.ac.kr, hjyang@jnu.ac.kr, and shkim@jnu.ac.kr

Abstract. This research solves the fundamental high-dimensional classification problem in machine learning. A classifier's performance may suffer as the number of attributes in the data is too large since there are fewer training samples available. Limiting the number of features through feature selection is one approach to overcoming this difficulty. Unlike earlier methods, ours proposes selecting features based on knowledge of how they will interact. Our approach employs the transfer learning with Residual Neural Network to extract the deep features first and select the optimal features by using principal component analysis (PCA) and T-distributed Stochastic Neighbor Embedding (t-SNE). Our approach is efficient and is able to feed the optimal features to the classifiers instead of feeding irrelevant information. Experiment Results on two high dimensional datasets shows the performance of our approach in the form of reduction of time and overall cost.

Keywords: Transfer Learning · Residual Neural Network · PCA and t-SNE.

1 Introduction

In recent years, it has become increasingly crucial to be able to read a person's emotional state. Human emotion recognition has garnered attention in several fields, including but not limited to human-computer(8), academia, and medical. Effective role of feelings in interpersonal communication is an impossibility. Emotions have a crucial role in regular human conversation. Sensors can learn about a person's mental state by listening to their voice and reading their facial expressions and body language. The dash et al. (2) states that nonverbal cues, such as voice tone and body language, make up 38% and 55%, respectively, of daily communication, whereas verbal cues account for only 7%.

Emotions can be read from the face, the tone of voice, and the body language of a person. Researchers have found that facial expressions are the most effective means of communicating feelings. Evidence of them can be presented in many forms, some of which are readily apparent to the naked eye and others of which are not.

Research on emotions draws from several fields, including psychology and computer science. In psychological words, it's a condition that influences one's way of thinking, feeling, and behaving, as well as one's level of contentment with life (9). In contrast, in the area of computer science, it can be identified in the form of visual, auditory, and textual data. It's not simple to extract feelings from any of these signals. Emotions, whether happy, negative, or neutral, are the primary means by which humans express themselves to one another. It's commonly recognized that words like "cheerful," "happy," and "excited" are used to portray positive emotions, while words like "hate," "anger," "fear," "depression," and "sad" are used to indicate negative ones. Social media platforms like Facebook, Instagram, and others have become the primary means by which people share information and communicate their emotions (10). They provide a wide range of outlets for expressing inner states.

The constraint of dimensionality can be overcome in pattern recognition by feature selection. Feature selection refers to the steps used to narrow down a large feature set to a manageable subset (1) (2). Features that are particular to a given class and do not overlap with other features make up the best feature set. For high-dimensional data in particular, feature selection is a crucial preprocessing step in machine learning since it reduces the cost of data gathering, aids in the discovery of crucial traits, and improves classification accuracy. In the recent century, high-dimensional data has proliferated, making feature selection more important than ever.

Filter methods, wrapper methods, and embedding methods are the three main types of feature selection strategies used in supervised learning. Methodologies for filtering data (3) (4) assess the value of a subset based on key features, such as information-based measures, distances, or statistical data. These techniques are particularly effective since they do not rely on a learning classifier to select the relevant characteristics. Consequently, it is a feature selection approach for high-dimensional data sets, although the outcomes are subpar. In contrast, wrapper techniques assess the usefulness of a given classifier to measure the importance of a narrowed subset and during search (5) (6). Wrapper approaches for high-dimensional data are time-consuming, but they outperform filter techniques for the same amount of extracted features(7). Each of these approaches, however, has its own limitations. Compared to the wrapper method, which relies on a centralized mechanism for evaluating and selecting features, the filter method relies on decentralized, independent evaluation to arrive at its final decision (6). The embedded approaches (8) (9) interact with the specific structure of a classifier, like the support vector machine (SVM) classifier or even the decision tree classifier, to select a set of characteristics that will be useful for classification. Consequently, this approach can restrict just few classifiers.

2 Related Work

The development of deep learning has substantially enhanced the precision of facial emotion identification. To solve the difficulties of emotion recognition from

facial expressions, many new Convolutional Neural Network (CNN) models have been invented recently. It is a top network in its industry. The building blocks of a convolutional neural network (CNN) are convolutions, activation layers, and pooling layers. Computing terminal devices can now interpret the variations in human emotions to an extent, resulting in more diversity in human-computer communication (11), thanks to the development of Artificial Intelligence technologies like pattern recognition and computer vision. The primary goal of face recognition (FER) is to assign a certain emotional state to a particular facial expression. Feelings can be identified by parsing a face image for recognizable traits and using those elements to identify the subject's emotional state. Face photos require some processing before being fed into a convolutional neural network (CNN) or other machine learning classifier. Existing techniques include the viola-jones algorithm(15), the histogram of gradients (14), the histogram equalization (13), the linear discriminant analysis(12), the histogram of discrete wavelets (12), etc.

Manual feature extraction excels at identification in controlled lab settings but struggles in real-world settings with factors like occlusion and lighting. Recently, there has been a lot of interest in using deep convolutional neural networks for feature extraction(16), which has improved the accuracy of face emotion identification. With its novel strategy to Deep Neural Network optimization, the Deep Residual Network(17)(Deep ResNet) has made significant strides in the field of image recognition. The two-stage classical learning technique used in earlier study on emotion recognition has been abandoned. In the initial phase, we employ image processing methods for feature extraction. Conversely, in the second phase, we used a classic machine learning classifier like Support Vector Machine (SVM) to identify feelings. Weighted random forest (WRF)(18) is one way FER has utilized to glean the most important features of image compositions. Hasani and Mahoor(19) employed a novel network called ResNet-LSTM, which integrates lower highlights to LSTMs, to capture Spatio-temporal data. Because of its improved feature extraction capabilities, the deep learning network has become the most preferred approach to FER. In the wavelet domain, Nigam et al.(14) provided a four-step procedure for efficient FER based on the histogram of oriented gradients (HOG) (face processing, domain transformation, feature extraction and expression recognition). The scientists used a tree-based multi-class SVM to categorize the HOG features obtained by the discrete wavelet transformation and then applied those findings to facial emotion identification. The CK+, JAFFE, and Yale datasets were used for training and testing. Three datasets have been examined, with the results showing an accuracy of 90%, 71.43%, and 75% in the test set.

After extensive research into the Facial Expression Recognition issue, Minaee et al. presented an Attentional Convolutional Neural Network (20) as an alternative to simply adding more layers/neurons. In addition, they proposed using a visualization tool that, using the classifier's output, can zero in on crucial regions of the face for discerning a variety of emotions. An integral aspect of their design is a spatial transformer network that performs an efficient transformation

to encase the input and provide an appropriately transformed output. For the 7 classes they were tasked with classifying, they used the FER2013 dataset and achieved an accuracy rate of 70.02 percent.

The authors utilized the Residual Masking Network(21) to zero in on deep architecture via the attention mechanism. To improve feature maps, they trained a segmented network to zero in on only the data it needed to make an informed call. They split their work into two sections: the residual masking block, which includes a residual layer, and also the ensemble approach for the conjunction with seven separate CNNs. Their final accuracy on the FER2013 dataset test set was 74.14%. Using a feature extraction network and a pre-trained model, Pu and Zhu(22) created a FER framework. Using a supervised learning technique called residual block optical flow, we may extract useful features.

Inception is used as a classifier for its innovative design. In tests on CK+ and FER2013 datasets, they were able to improve accuracy to 95.74 and 73.11 percent, respectively. Chowanda(23) developed a separable CNN to address the issue of how much computing power is needed to train and analyses CNNs for emotional recognition. As part of the experiment, we compared four different kinds of networks against one another. Networks both with and without modular separation, with and without flattening and completely connected layers, and with and without resorting to global average pooling. They got an accuracy of 99.4 percent on the CK+ dataset with their proposed architecture, and it was faster and had fewer parameters. In our method we used a novel method to choose the optimal number of feature to feed the classifier instead of feeding all the information for this purpose we use PCA and t-SNE. Our method showed efficient performance on two datasets. Remaining paper is consist of proposed method, experiment & results and conclusion.

3 Proposed Method

There are four layers on masking network, each Residual Masking Block includes a Residual Layer and a Masking Block that performs its function on features of varying sizes. After being passed through a 3x3 convolutional layer with stride 2, an input image of size 224x224 will be passed through a 2x2 max-pooling layer, reducing its spatial size to 56x56. We optimised it with triplet loss function and feed the deep features to PCA and t-SNE for dimension reduction. By feeding we obtained the optimal deep features with only the most impact 20 dimensions remained. The following four Residual Masking Blocks turn the feature maps derived from the preceding pooling layer into feature maps with four different spatial sizes: 56 by 56, 28 by 28, 14 by 14, and 7 by 7. At the very end, the network employs a pooling layer to average the inputs and a softmax layer with 7 inputs to generate outputs that map to 7 different expressions (6 emotions and one neutral state). The proposed method is showed in Figure 1

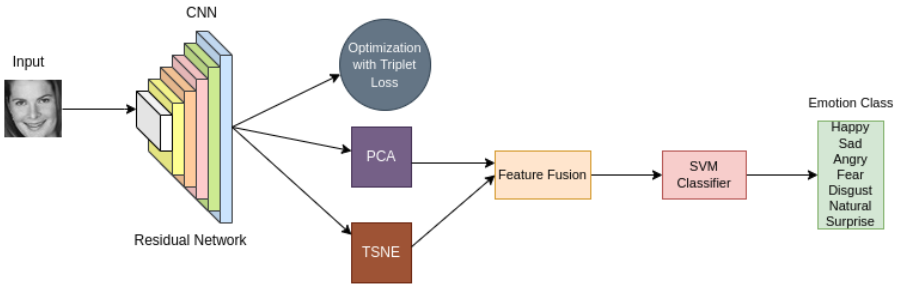


Fig. 1: Proposed Method

4 Experiment and Resuts

4.1 Dataset

In our experiment three public dataset CK+48, JAFFE and FER2013 are used. At ICML 2013's Challenges in Representation Learning, the first widely used dataset, FER2013(24), was unveiled. As can be seen in Figure 2, there are a total of 35887 greyscale (48x48) photos inside this collection. To train the model, we used 28709 photos; to validate it, we used another 3850; and to test it, we used 3589. Google's image search API gathers all of these pictures and assigns labels for anger, disgust, fear, happiness, sadness, surprise, and neutral. This data set is commonly used for benchmarking various FER approaches that involve deep learning.

Dataset	Type	#Sample	#Feature	#Classes
CK+48	Face image	981	20	7
JAFFE	Face image	213	20	7

Table 1: Number of Features, Samples, and Class in Each DataSet

4.2 Experimental Setup

In order to train with pre-trained models from ImageNet, the original training images are scaled to 224 x 224 and converted to RGB. In addition, over-fitting is avoided by augmenting training photos. Rotating by a factor of(25) is one of the augmentation techniques, along with a left-right flip. For each experiment, if the validation accuracy does not improve by at least eight steps after 50 epochs, the experiment is terminated. Scheduler reduces learning rate by a factor of 10 if validation accuracy does not improve over five consecutive epochs while using

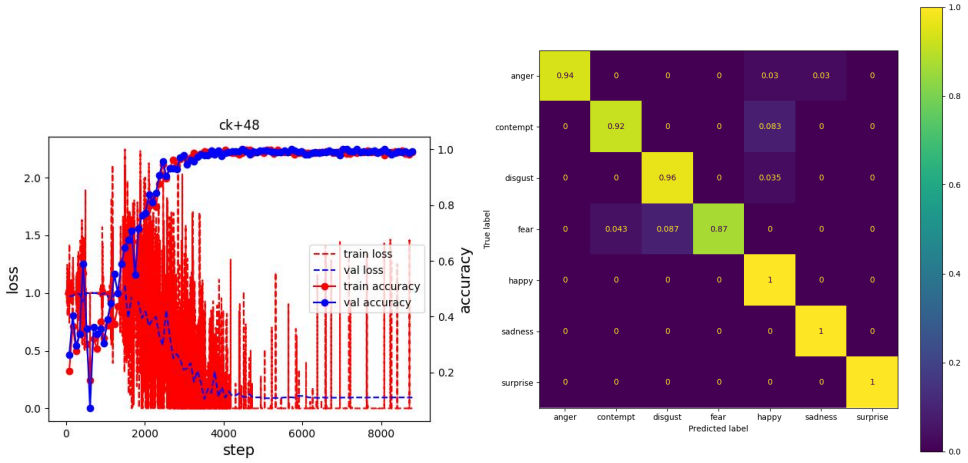
a batch size of 48 and an initial learning rate of 0.0001. Network-agnostic experiments are run with the same hyperparameters, preprocessing, augmentation, and evaluation metrics as one another Pytorch is used for the experiments, and the graphics card used is a RTX 3090.

Processing time in the actual application is tested using a desktop computer equipped with a AMD® Ryzen 7 2700x eight-core processor × 16 CPU, a Graphics Processing Unit (GPU) GTX 1050Ti, and 24GB of RAM. The suggested network can handle 100 face-containing frames per second with the current setup. This finding gives us confidence in the practicality of our application in real time.

	CK+48
Proposed method	99.66%
Deep Features + SVM classifier	98.83%

Table 2: Performance comparison between our proposed pipeline with previous works on CK+48 dataset.

We conduct two experiments on CK+48 to assess the effectiveness of the proposed method with one experiment using the proposed method and the other feeding deep features to the SVM classifier directly. The dataset is splited into training and validation subsets with 7/3 ratio. Table 2 shows that the proposed method increases accuracy on CK+48 with almost 1%. We also provide the training curve and confusion matrix in Figure 2. Following the confusion matrix, the worst performance is in the "fear" class and all the other classes achieve almost perfect performance.



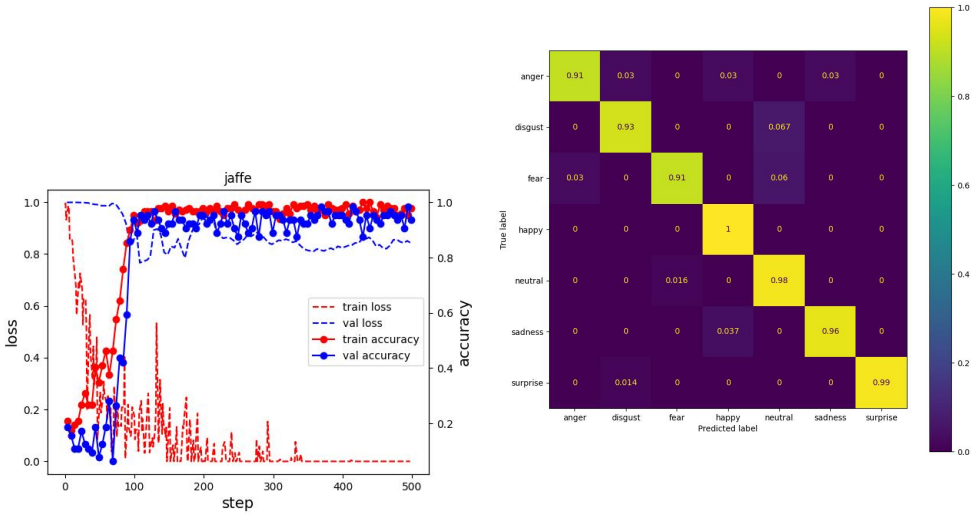
(a) Training curve (b) Confusion matrix

Fig. 2: Training curve and confusion matrix on CK+48

	JAFFE
Proposed method	98.79%
Minaee and Abdolrashidi (2019)	92.80%
Khaireddin and Chen (2021)	73.28%
Aouayeb et al. (2021)	94.83%
Boughida et al. (2022)	96.30%
Shaik and Cherukuri (2022)	97.46%

Table 3: Performance comparison between our proposed pipeline and previous works on the JAFFE dataset.

We also experiment on JAFFE dataset to compare our proposed method with other previous works. We split the dataset into training and validation subset with ratio of 7/3. The Table 3 shows that our proposed method outperforms all the previous works and achieves the SoTA result on JAFFE dataset with 98.79%. The confusion matrix in Figure 3 also shows the near-perfect performance on all emotion classes in JAFFE dataset.



(a) Training curve (b) Confusion matrix

Fig. 3: Training curve and confusion matrix on JAFFE

To further evaluate our proposed method, we perform the proposed method on a larger and more complex dataset than previous ones and show that the proposed method can work well on a small amount of the original FER2013 dataset. We conduct experiments on 5%, 10% and 15% of FER2013 datasets. The experiment results are showed in Table 4. The results shows that our proposed method can achieve the comparable results with other previous SoTA works.

	5% FER2013	10% FER2013	15% FER2013
Proposed method	53.45%	57.24%	61.76%
Barsoum et al. (2016)	53.37%	57.43%	60.11%
Khaireddin and Chen (2021)	53.56%	58.43%	61.32%
Pham et al. (2021)	54.12%	60.60%	61.60%

Table 4: Performance comparison between our proposed pipeline and previous works on the FER2013 dataset.

5 Conclusion

This research improves upon existing methods for recognizing facial expressions by introducing a new feature selection idea, which is implemented by using transfer learning with a Residual Neural Network. In this Residual Neural Network, different Masking Blocks are applied throughout Residual Layers to improve the optimal features selection method by using the PCA and t-SNE. The experimental results obtained using the CK+48 and JAFFE dataset proved that the proposed methods outperformed the state-of-the-art results and the most popular classification algorithms. The proposed method will be further developed with the goal of assessing the model's generalization using the largest existing classification dataset. To further boost network performance in vision tasks like classification and detection, we will investigate varying network parameters and attempt to reduce the number of model parameters required for these tasks. We intend to construct a whole system and put it through its paces in a public rehearsal setting.

Acknowledgements This research was supported by the Bio Medical Technology Development program of the National Research Foundation(NRF) funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2018R1D1A3B05049058 NRF-2020R1A4A1019191).

Bibliography

[1] Kwak, Nojun, and Chong-Ho Choi. "Input feature selection for classification problems." IEEE transactions on neural networks 13, no. 1 (2002).
[2] Dash, Manoranjan, and Huan Liu. "Feature selection for classification." Intelligent data analysis 1, no. 1-4 (1997).
[3] Bommert, Andrea, Xudong Sun, Bernd Bischl, Jorg Rahnenfuhrer, and Michel Lang. "Benchmark for filter methods for feature selection in high-dimensional classification data." Computational Statistics Data Analysis 143 (2020).
[4] Nguyen, Hoai Bach, Bing Xue, Ivy Liu, and Mengjie Zhang. "Filter based backward elimination in wrapper based PSO for feature selection in classification." In 2014 IEEE congress on evolutionary computation (CEC), pp. 3111-3118. IEEE, 2014.
[5] Mustaqeem, Anam, Syed Muhammad Anwar, Muhammad Majid, and Abdul Rashid Khan. "Wrapper method for feature selection to classify cardiac arrhythmia." In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3656-3659. IEEE, 2017.
[6] Zhang, Jixiong, Yanmei Xiong, and Shungeng Min. "A new hybrid filter/wrapper algorithm for feature selection in classification." Analytica chimica acta 1080 (2019).

[7] Labani, Mahdieh, Parham Moradi, Fardin Ahmadizar, and Mahdi Jalili. "A novel multivariate filter method for feature selection in text classification problems." Engineering Applications of Artificial Intelligence 70 (2018).
[8] R. Cowie et al., "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32–80, 2001.
[9] B. Parkinson and A. S. R. Manstead, "Current Emotion Research in Social Psychology: Thinking About Emotions and Other People," Emotion Review, vol. 7, no. 4, pp. 371–380, Jul. 2015.
[10] S. F. Waterloo, S. E. Baumgartner, J. Peter, and P. M. Valkenburg, "Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp," New Media Society, vol. 20, no. 5, pp. 1813–1831, May 2017.
[11] V. R. LeBlanc, M. M. McConnell, and S. D. Monteiro, "Predictable chaos: a review of the effects of emotions on attention, memory and decision making," Advances in Health Sciences Education, vol. 20, no. 1, pp. 265–282, Jun. 2014.
[12] Y. Zhu, "Research on the Human-Computer Interaction Design in Mobile Phones," 2020 International Conference on Computing and Data Science (CDS), Aug. 2020.
[13] N. Chervyakov, P. Lyakhov, D. Kaplun, D. Butusov, and N. Nagornov, "Analysis of the Quantization Noise in Discrete Wavelet Transform Filters for Image Processing," Electronics, vol. 7, no. 8, p. 135, Aug. 2018.
[14] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition," Procedia Computer Science, vol. 116, pp. 523–529, Jan. 2017.
[15] S. Nigam, R. Singh, and A. K. Misra, "Efficient facial expression recognition using histogram of oriented gradients in wavelet domain," Multimedia Tools and Applications, vol. 77, no. 21, pp. 28725–28747, May 2018.
[16] N. Deshpande and S. Ravishankar, "Face Detection and Recognition using Viola-Jones algorithm and Fusion of PCA and ANN," vol. 10, no. 5, pp. 1173–1189, 2017.
[17] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," openaccess.thecvf.com, 2016.
[19] Z.-S. Liu, W.-C. Siu, and J.-J. Huang, "Image super-resolution via weighted random forest," IEEE Xplore, Mar. 01, 2017.
[20] B. Hasani and M. H. Mahoor, "Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields," IEEE Xplore, May 01, 2017.
[21] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," Sensors, vol. 21, no. 9, p. 3046, Apr. 2021.
[22] L. Pham, T. H. Vu, and T. A. Tran, "Facial Expression Recognition Using Residual Masking Network," IEEE Xplore, Jan. 01, 2021.

[23] L. Pu and L. Zhu, "Differential Residual Learning for Facial Expression Recognition," 2021 The 5th International Conference on Machine Learning and Soft Computing, Jan. 2021.
[24] Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," Neural Information Processing, pp. 117–124, 2013.
[25] Y. S. Teo et al., "Benchmarking quantum tomography completeness and fidelity with machine learning," New Journal of Physics, vol. 23, no. 10, p. 103021, Oct. 2021.
[26] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," Proceedings of the 18th ACM International Conference on Multimodal Interaction, Oct. 2016.
[27] Wang, K., Peng, X., Yang, J., Meng, D. & Qiao, Y. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *IEEE Transactions On Image Processing*. **29** pp. 4057-4069 (2019)
[28] Siqueira, H., Magg, S. & Wermter, S. Efficient Facial Feature Learning with Wide Ensemble-based Convolutional Neural Networks. *ArXiv*. **abs/2001.06338** (2020)
[29] Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K. & Qiao, Y. Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. *2019 International Conference On Multimodal Interaction*. (2019)
[30] Happy, S. & Routray, A. Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions On Affective Computing*. **6** pp. 1-12 (2015)
[31] Fard, A. & Mahoor, M. Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild. *IEEE Access*. pp. 1-1 (2022)
[32] Farzaneh, A. & Qi, X. Facial Expression Recognition in the Wild via Deep Attentive Center Loss. *Proceedings Of The IEEE/CVF Winter Conference On Applications Of Computer Vision (WACV)*. pp. 2402-2411 (2021,1)
[33] Khaireddin, Y. & Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. *CoRR*. **abs/2105.03588** (2021), https://arxiv.org/abs/2105.03588
[34] Savchenko, A., Savchenko, L. & Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions On Affective Computing*. **13**, 2132-2143 (2022)
[35] Bodapati, J., Srilakshmi, U. & Veeranjaneyulu, N. FERNet: a deep CNN architecture for facial expression recognition in the wild. *Journal Of The Institution Of Engineers (India): Series B*. **103**, 439-448 (2022)
[36] Oguine, O., Oguine, K., Bisallah, H. & Ofuani, D. Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction. *ArXiv Preprint ArXiv:2206.09509*. (2022)
[37] Shaik, N. & Cherukuri, T. Visual attention based composite dense neural network for facial expression recognition. *Journal Of Ambient Intelligence And Humanized Computing*. pp. 1-14 (2022)

[38] Boughida, A., Kouahla, M. & Lafifi, Y. A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. *Evolving Systems*. **13**, 331-345 (2022)

[39] Qi, Y., Zhou, C. & Chen, Y. NA-Resnet: neighbor block and optimized attention module for global-local feature extraction in facial expression recognition. *Multimedia Tools And Applications*. pp. 1-19 (2022)

[40] Abdulsattar, N. & Hussain, M. Facial Expression Recognition using Transfer Learning and Fine-tuning Strategies: A Comparative Study. *2022 International Conference On Computer Science And Software Engineering (CSASE)*. pp. 101-106 (2022)

[41] Minaee, S. & Abdolrashidi, A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors (Basel, Switzerland)*. **21** (2019)

[42] Aouayeb, M., Hamidouche, W., Soladié, C., Kpalma, K. & Séguier, R. Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition. *ArXiv*. **abs/2107.03107** (2021)

[43] Barsoum, E., Zhang, C., Canton-Ferrer, C. & Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. *Proceedings Of The 18th ACM International Conference On Multimodal Interaction*. (2016)

[44] Khaireddin, Y. & Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. *ArXiv*. **abs/2105.03588** (2021)

[45] Pham, L., Vu, T. & Tran, T. Facial Expression Recognition Using Residual Masking Network. *2020 25th International Conference On Pattern Recognition (ICPR)*. pp. 4513-4519 (2021)