Face Emotion Recognition Using Convolutional Neural Network

Abstract

Emotion is a significant factor in human lives. It affects human decision-making and reasoning. Facial expressions are one of the main sources in classifying human emotion. In this study, we proposed a CNN-based model to classify four emotions – *angry, happy, sad, surprise* with an additional *neutral* class. The model was trained on 100x100 pixel face images with an applied Sobel filter for edge detection. The proposed CNN model resulted in 77.3% accuracy outperforming the benchmark model with 70.1% accuracy. In addition, the model performed better in classifying *angry, happy, sad, surprise* emotions.

1. Introduction

Emotion is an important part of our daily lives. It plays a significant role in human perception, attention, memory, reasoning, etc. Emotion is a complex process involving response in different mediums like physiological system, facial and vocal expressions, and cognition (Kassam KS, Mendes WB, 2013). It happens involuntarily and affects how the human mind and body behave. Paul Ekman suggests that emotion can be characterized into six basic classes – happiness, sadness, fear, disgust, anger, and surprise (Ekman P., 1992). Psychologists propose different emotion categories but there has been a substantial agreement on these six basic emotions (Jerritta et al., 2011). According to Tomkins, emotion is the foundation of human motivation, and that human faces are the main area of emotion (Tomkins, 2014). Humans are skillful in recognizing and discriminating emotion through facial expressions. In addition, certain emotions are more visible on specific parts of the face. According to Wegrzyn et al. (2017), emotions like disgust, surprise, and happiness are more noticeable on the lower part of the face while anger, neutral, fear, and sad are more noticeable on the upper part (Wegrzyn et al., 2017). Therefore, the face is a reliable basis in determining the current emotional state of a person. Facial expression recognition (FER) can be applied to various areas like psychology and social interaction. In robotics, emotion recognition improves human-to-computer interaction (HCI) (Giorgana, G., & Ploeger, P. G., 2012). Interest in FER has been increasingly aligned with the rapid development in the field of artificial intelligence. Conventional FER approaches classify expression using features extracted from facial images (Samadiani, et al., 2019). Popular methods like Hidden Markov Model (HMM), Support Vector Machine (SVM), AdaBoost, and Artificial Neural Networks (ANN) are used in classifying these facial features (Takalkar et al., 2017). In this paper, we applied Convolutional Neural Network (CNN), a deep learning algorithm commonly used in image classification, to train our model using the FER2013 dataset (Kaggle, 2020).

2. Related Studies

The selection of image features is critical in developing a classification model. The classifier relies heavily on relevant and distinguishable image information (Kumar, G., & Bhatia, P. K., 2014). Generally, face-based recognition uses extracted facial features and applies algorithms for developing a classification model. The method proposed in (Roomi, S. M. M., et al., 2011) uses the Viola-Jones algorithm in detecting faces from images then applying Sobel edge detection image-processing in extracting facial features. In the research of (Asad, M., et al, 2017), they also utilized Viola-Jones object detection for detecting different human facial features called Haar-like features to train a model in predicting seven essential facial expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise). Another approach was proposed in (Khan, F., 2018) using facial landmarks as a basis in recognizing facial expressions. They used Euclidean distances to formulate these facial landmarks before feeding them into a Multi-Layer Perceptron (MLP) neural network to classify different emotions. In the study of (Pitaloka, D. A., et al., 2017), the method involves cropping the region of interest (ROI) of the face using the Haar algorithm then applying pre-processing techniques – resizing, adding noise, and data normalization.

The majority of research in emotion recognition uses different approaches in extracting facial features but the Haar Cascade classifier is mostly used as an initial process in feature extraction.

Several studies are trying to understand how humans perceive emotions on static faces. One of the main approaches is using the convolutional neural network (CNN), a deep learning method to classify human emotions based on facial expressions in images. In the research of (Rakovitsky, A., & Knott, J., 2020), they modified the CNN architecture of ResNet50 to match their objective in predicting emotions on grayscale images of art paintings subjects. They added a dropout layer, a fully connected layer, and seven output nodes corresponding to the following emotions – Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral which resulted in a 52.9% classification accuracy. In another conducted as an entry on emotion recognition competition EmotiW 2017 (Knyazev et. al., 2017), the proposed model is the combination of four CNN architectures – VGG-Face which is pre-trained with around three million images along with three other proprietary networks achieved an overall accuracy of 60.03% on emotion classification. In (Khaireddin, Y., & Chen, Z., 2021), the model consists of four convolutional layers each having a Rectified Linear Unit (ReLU) activation function. It classifies Ekman's six basic emotions with neutral expression from a 40x40 human face image. The research reports an accuracy of 62.44 % with the CNN model using the FER2013 dataset.

According to (Khaireddin, Y., & Chen, Z., 2021), several CNN models achieved a low classification accuracy ranging from 65% to 72.7% using the FER2013 dataset. The reason for this was elaborated in (Giannopoulos et. al., 2017) where they discussed the difficulty of classification due to certain conditions like facial pose, lighting, shadows, and orientations with the highlight on non-uniformity of human faces.

3. Methodology

I. Dataset

In this research, we trained the emotion classification model using the Facial Emotion Recognition 2013 (FER2013) image dataset. The images are in 48x48 pixels grayscale format and are collected using a Google image search of each emotion. Due to a high imbalance in the dataset, we only focus on classifying four emotions – angry, happy, sad, and surprise with additional neutral. Other image samples from different sources like Extended Cohn-Kanade (CK+) and AffectNet were also gathered to balance the dataset.

Fig 1. Sample image per emotion.

The dataset consists of 35,770 facial images corresponding to four basic human emotions with an additional neutral class. The happiness emotion has 7165 images, sadness has 7208 images, anger has 7145 images, surprise has 7089 images, and neutral has 7160 images. In addition, the dataset is divided into 28,709 training images and 7,178 test images. The dataset was split into train-validation-test in the ratio of 60:25:15.

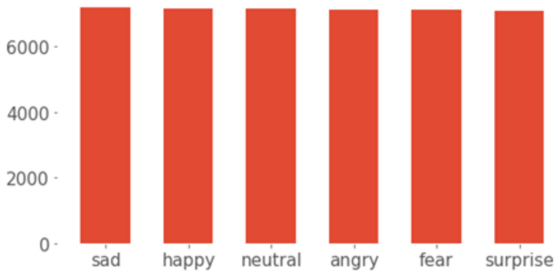


Fig 2. Image distribution per emotion.

II. Preprocessing

In extracting facial features, gaussian blur was applied to the image to minimize noises in the image before using an edge detection filter called the Sobel operator from the OpenCV library. The Sobel operator is based on the convolution of the image in the horizontal and vertical directions. It calculates the approximate gradient of an image using a 3x3 kernel (OpenCV).

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Fig 3. Sobel operator masks.

The result of two 3x3 kernels denoted and can be combined to find the absolute magnitude of the gradient at each point on an image using the equation:

$$G = \sqrt{G_x^2 + G_y^2}$$

Since Sobel is sensitive to image noise (Beeran et. al., 2014), gaussian blur was applied to the image before applying the edge detection filter.

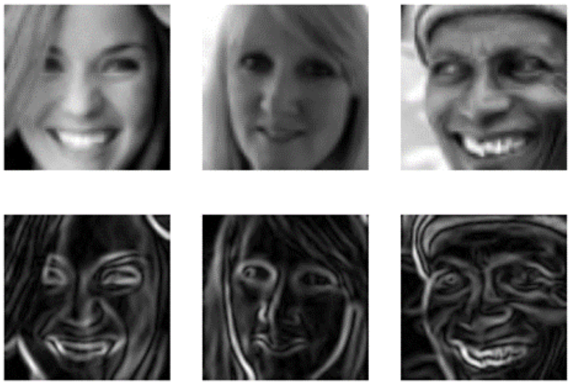


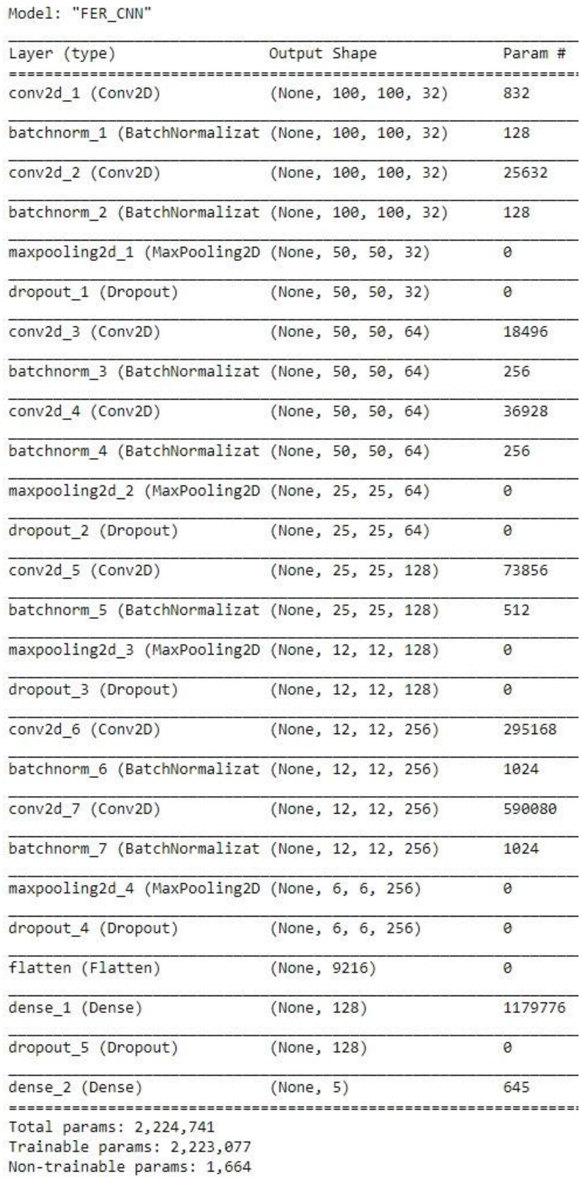
Fig 4. Results of Sobel edge detection

Traditional image data augmentation based on the TensorFlow library was utilized to further increase the training dataset and introduce different image conditions to the classification model. The augmentation involves rotating the image by ± 10 degrees, shifting the width by ± 20 percent and height by ± 10 percent, zooming the image by ± 10 percent, shearing transformation by 10 degrees, and flipping images horizontally. These techniques were applied randomly on the training set during the model training. The image was resized to 100x100 pixels and each pixel value was divided into 255 to normalize the data before inserting it into the CNN model.

III. Model

We aim to find the accuracy of a Convolutional Neural Network in classifying emotion through facial expression. The proposed model is a deep convolutional neural network that classifies five emotions on the face images dataset. The network consists of seven convolution layers each having an Exponential Linear Unit (ELU) activation function followed by a Batch Normalization and Max Pooling layer with a 2 x 2 pool size. A dense layer with 128 filters was added before the output layer. Dropouts were also introduced to the model at regular intervals to avoid overfitting.

We modified the model input shape to 100x100 instead of 48x48 that was used on the base model.



```
Model: "FER_CNN"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 100, 100, 32)      832
_________________________________________________________________
batchnorm_1 (BatchNormalizat (None, 100, 100, 32)      128
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 100, 100, 32)      25632
_________________________________________________________________
batchnorm_2 (BatchNormalizat (None, 100, 100, 32)      128
_________________________________________________________________
maxpooling2d_1 (MaxPooling2D (None, 50, 50, 32)        0
_________________________________________________________________
dropout_1 (Dropout)          (None, 50, 50, 32)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 50, 50, 64)        18496
_________________________________________________________________
batchnorm_3 (BatchNormalizat (None, 50, 50, 64)        256
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 50, 50, 64)        36928
_________________________________________________________________
batchnorm_4 (BatchNormalizat (None, 50, 50, 64)        256
_________________________________________________________________
maxpooling2d_2 (MaxPooling2D (None, 25, 25, 64)        0
_________________________________________________________________
dropout_2 (Dropout)          (None, 25, 25, 64)        0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 25, 25, 128)       73856
_________________________________________________________________
batchnorm_5 (BatchNormalizat (None, 25, 25, 128)       512
_________________________________________________________________
maxpooling2d_3 (MaxPooling2D (None, 12, 12, 128)       0
_________________________________________________________________
dropout_3 (Dropout)          (None, 12, 12, 128)       0
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 12, 12, 256)       295168
_________________________________________________________________
batchnorm_6 (BatchNormalizat (None, 12, 12, 256)       1024
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 12, 12, 256)       590080
_________________________________________________________________
batchnorm_7 (BatchNormalizat (None, 12, 12, 256)       1024
_________________________________________________________________
maxpooling2d_4 (MaxPooling2D (None, 6, 6, 256)         0
_________________________________________________________________
dropout_4 (Dropout)          (None, 6, 6, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 9216)              0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               1179776
_________________________________________________________________
dropout_5 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 5)                 645
=================================================================
Total params: 2,224,741
Trainable params: 2,223,077
Non-trainable params: 1,664
```

Fig 5. Proposed CNN architecture

Many facial emotion recognition model benchmarks were trained using transfer learning. However, since our study does not include a pretrained model, we will be using the CNN model developed in (Sharma, 2020) as our benchmark in model evaluation. The benchmark model was a submission on Kaggle and was used to train on the fer2013 dataset.

4. Results

In order to evaluate our model, Stratified K-fold validation along with Paired T-Test was used to determine if there are significant differences in proposed and benchmark model results.

Fold	Benchmark	Proposed
K1	71.6	77.2
K2	69.1	77.4
K3	71.5	77.5
K4	71.8	77.4
K5	66.5	77.1
Average accuracy	70.1	77.3
p-value		0.0019

Table 1. Stratified cross-fold validation of prediction accuracy

Table 1 shows that the benchmark model with 70.1% average accuracy was outperformed by the proposed model with 77.3% average accuracy. The results on cross-fold validation accuracy were used to perform a T-test. The result shows a p-value of 0.0019 which means that there is a significant difference between the proposed model and the benchmark model.

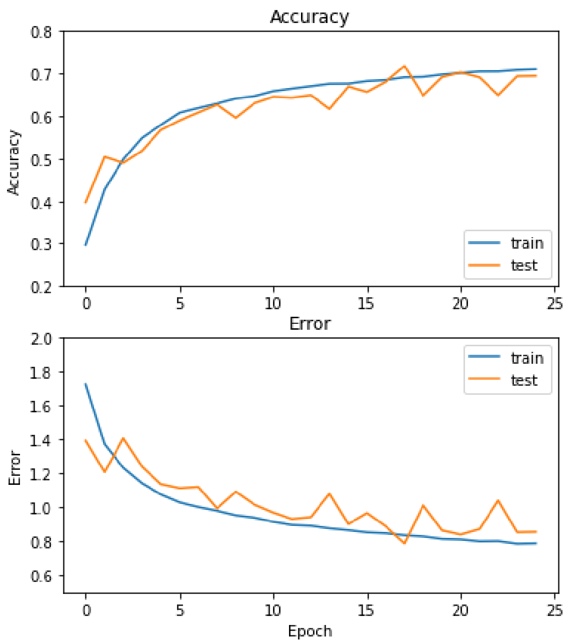


Fig 6. Accuracy and error of benchmark model

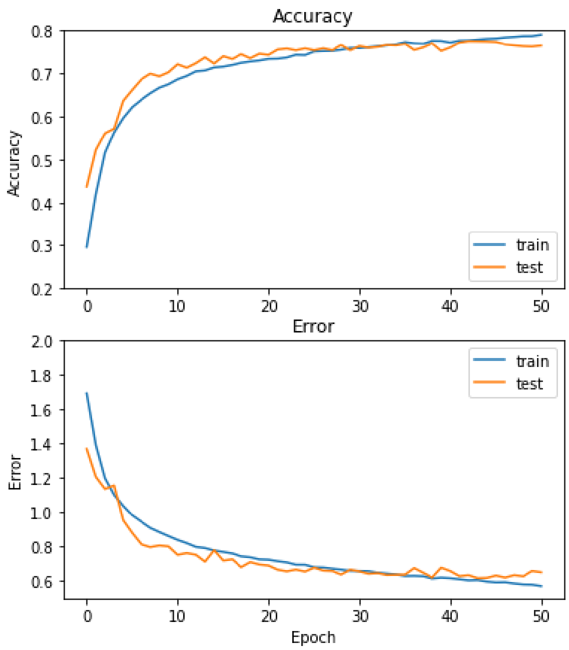


Fig 7. Accuracy and error of the proposed model

In Fig 6 and 7, we can observe that there is a minimal fluctuation in the test loss and accuracy of the benchmark model which is a sign that the model experiences minimal overfitting. On the other hand, the performance of the proposed model shows that there is minimal to no overfitting. In addition, the benchmark model reached its peak accuracy of 70% on 21 epochs while the proposed model reached its peak accuracy of 77% on 42 epochs. Although the proposed model took more epochs, it still outperformed the benchmark model in terms of accuracy.

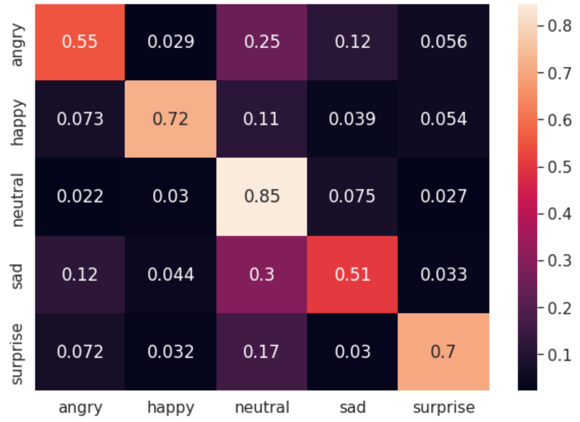


Fig 8. Confusion matrix of the benchmark model

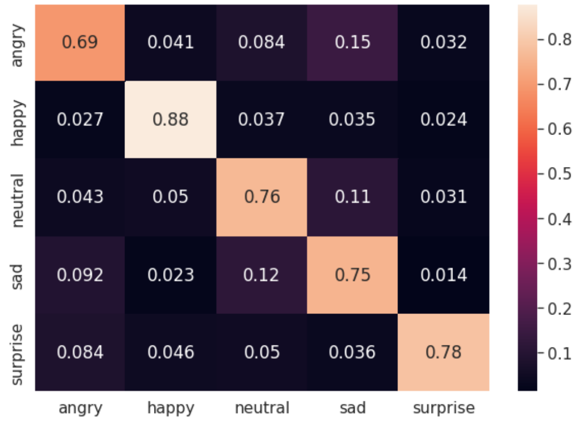


Fig 9. Confusion matrix of the proposed model

Since we are dealing with multi-class classification, we trained each model on 5,366 sample images and plotted the confusion matrix of both benchmark and proposed model in order to visualize the accuracy of the models on each emotion. The proposed model outperformed the benchmark model in classifying the emotion *angry, happy, sad, surprise* with an exception to the *neutral* class.

5. Conclusion

In this paper, we presented a CNN-based model to classify emotion on face images. The proposed model accuracy was recorded at 77.3% outperforming the benchmark model that has 70.1% accuracy. In classifying each emotion, the proposed model performed better in classifying *angry, happy, sad, and surprise* emotions while the benchmark model performed better in classifying *neutral*.

The proposed model can be used in evaluating the audience during presentations. It can also be applied in an e-learning environment to give evaluations on student responses during discussions.

For future work, we would like to improve the model by classifying more emotions. This includes the primary and secondary emotions. Additional improvements can be done to improve the training time and accuracy of the model.

References

Asad, M., Gilani, S. O., & Jamil, M. (2017). Emotion Detection through Facial Feature Recognition. International Journal of Multimedia and Ubiquitous Engineering, 12(11), 21–30. https://doi.org/10.14257/ijmue.2017.12.11.03

Beeran Kutty, S., Saaidin, S., Megat Yunus, P. N. A., & Abu Hassan, S. (2014). Evaluation of canny and Sobel operator for logo edge detection. 2014 International Symposium on Technology Management and Emerging Technologies. Published. https://doi.org/10.1109/istmet.2014.6936497

Ekman, P. (1992). Are there basic emotions? Psychological Review, 99(3), 550–553. https://doi.org/10.1037/0033-295x.99.3.550

FER-2013. (2020, July 19). Kaggle. https://www.kaggle.com/msambare/fer2013

Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2017). Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013. Advances in Hybridization of Intelligent Methods, 1–16. https://doi.org/10.1007/978-3-319-66790-4_1

Giorgana, G., & Ploeger, P. G. (2012). Facial Expression Recognition for Domestic Service Robots. Lecture Notes in Computer Science, 353–364. https://doi.org/10.1007/978-3-642-32060-6_30

Jerritta, S., Murugappan, M., Nagarajan, R., & Wan, K. (2011). Physiological signals based human emotion Recognition: a review. 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications. Published. https://doi.org/10.1109/cspa.2011.5759912

Kassam, K. S., & Mendes, W. B. (2013). The Effects of Measuring Emotion: Physiological Reactions to Emotional Situations Depend on whether Someone Is Asking. PLoS ONE, 8(6), e64959. https://doi.org/10.1371/journal.pone.0064959

Khaireddin, Y., & Chen, Z. (2021). Facial Emotion Recognition: State of the Art Performance on FER2013. ArXiv, abs/2105.03588. https://arxiv.org/ftp/arxiv/papers/2105/2105.03588.pdf

Khan, F. (2018, December 10). Facial Expression Recognition using Facial Landmark Detection and Feature Extraction via Neural Networks. ArXiv.Org. https://arxiv.org/abs/1812.04510

Knyazev, B., Shvetsov, R., Efremova, N. & Kuharenko, A. (2017). Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. https://arxiv.org/pdf/1711.04598.pdf

Kumar, G., & Bhatia, P. K. (2014). A Detailed Review of Feature Extraction in Image Processing Systems. 2014 Fourth International Conference on Advanced Computing & Communication Technologies. Published. https://doi.org/10.1109/acct.2014.74

OpenCV: Sobel Derivatives. (n.d.). OpenCV. https://docs.opencv.org/3.4/d2/d2c/tutorial_sobel_derivatives.html

Pitaloka, D. A., Wulandari, A., Basaruddin, T., & Liliana, D. Y. (2017). Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition. Procedia Computer Science, 116, 523–529. https://doi.org/10.1016/j.procs.2017.10.038

Pranav, E., Kamal, S., Satheesh Chandran, C., & Supriya, M. (2020). Facial Emotion Recognition Using Deep Convolutional Neural Network. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). https://doi.org/10.1109/icaccs48705.2020.9074302

Rakovitsky, A., & Knott, J. (2020). WikiArt Analysis Using Facial Emotion Analysis. Arielrakovitsky. https://www.arielrakovitsky.com/papers/wikiart.pdf

Roomi, S. M. M., Virasundarii, S., Selvamegala, S., Jeevanandham, S., & Hariharasudhan, D. (2011). Race Classification Based on Facial Features. 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics. Published. https://doi.org/10.1109/ncvpripg.2011.19

Samadiani, Huang, Cai, Luo, Chi, Xiang, & He. (2019). A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data. Sensors, 19(8), 1863. https://doi.org/10.3390/s19081863

Sharma, G. (2020, April 18). Facial Emotion Recognition. Kaggle.

https://www.kaggle.com/gauravsharma99/facial-emotion-recognition/notebook

Takalkar, M., Xu, M., Wu, Q., & Chaczko, Z. (2017). A survey: facial micro-expression recognition. Multimedia Tools and Applications, 77(15), 19301–19325. https://doi.org/10.1007/s11042-017-5317-2

Tomkins Institute » Affect Imagery Consciousness (Vol I-IV). (2014). The Tomkins Institute. https://www.tomkins.org/what-tomkins-said/books-and-articles/affect-imagery-consciousness-vol-i-iv/

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. PLOS ONE, 12(5), e0177239. https://doi.org/10.1371/journal.pone.0177239