

# Facial emotion recognition using deep learning detector and classifier

Ng Chin Kit, Chee-Pun Ooi, Wooi-Haw Tan, Yi-Fei Tan, Soon-Nyeen Cheong

Faculty of Engineering, Multimedia University, Selangor, Malaysia

## Article Info

### Article history:

Received Feb 15, 2022

Revised Sep 1, 2022

Accepted Sep 2, 2022

### Keywords:

Convolutional neural network

Deep learning

Facial alignment

Facial emotion recognition

Facial landmark

## ABSTRACT

Numerous research works have been put forward over the years to advance the field of facial expression recognition which until today, is still considered a challenging task. The selection of image color space and the use of facial alignment as preprocessing steps may collectively pose a significant impact on the accuracy and computational cost of facial emotion recognition, which is crucial to optimize the speed-accuracy trade-off. This paper proposed a deep learning-based facial emotion recognition pipeline that can be used to predict the emotion of detected face regions in video sequences. Five well-known state-of-the-art convolutional neural network architectures are used for training the emotion classifier to identify the network architecture which gives the best speed-accuracy trade-off. Two distinct facial emotion training datasets are prepared to investigate the effect of image color space and facial alignment on the performance of facial emotion recognition. Experimental results show that training a facial expression recognition model with grayscale-aligned facial images is preferable as it offers better recognition rates with lower detection latency. The lightweight MobileNet\_v1 is identified as the best-performing model with WM=0.75 and RM=160 as its hyper-parameters, achieving an overall accuracy of 86.42% on the testing video dataset.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Chee-Pun Ooi

Faculty of Engineering, Multimedia University

63100 Cyberjaya, Selangor, Malaysia

Email: cpooi@mmu.edu.my

## 1. INTRODUCTION

Nonverbal components convey two-thirds of human communication [1]. Facial expression is one of the most effective non-verbal channels that helps people understand the internal emotions and intentions of others. Research on facial expression understanding has gathered significant attention over the years in both the research and industrial communities due to its great potential in many important technological applications such as computer animations, human-computer interaction, fatigue measurement, and crowd analytics. There are six basic emotional expressions that are considered universal among human beings regardless of different cultures and ethnicities, namely anger, disgust, fear, happiness, sadness, and surprise [2]. Existing facial emotion recognition techniques mainly focus on classifying the six basic expressions together with an additional “neutral” class. One popular approach is to follow the facial action coding system (FACS) [3] which estimates facial emotions based on the activation of different sets of action units (AU). There are also dimensional approaches that treat facial expressions as regression in the Arousal-Valence space to represent different emotional states [4].

In general, existing facial expression recognition (FER) algorithms are made up of three main modules, namely face/landmark detection, facial feature extraction, and facial expression classification. In this

context, FER algorithms can be divided into two groups depending on the nature of extracted facial features, either handcrafted or generated through the feature-learning process of deep neural networks. In conventional FER algorithms, handcrafted facial features are extracted from the detected face region, and pretrained facial expression classifier such as AdaBoost, k-nearest neighbor, and support vector machine (SVM) is used to estimate the emotions based on the extracted features. The handcrafted features can be geometric features (i.e., a feature vector is constructed based on the relationship between facial components), appearance features (i.e., a feature vector is formed by extracting image features from global face region or specific face regions with higher levels of importance for FER), or a combination of both. In [5], the position and angle of 52 facial landmark points are used as geometric features to construct the feature vector for the training classifier. The angle and Euclidean distance between each pair of landmark points are calculated for each frame and subtracted from the corresponding angle and distance in the first frame of a video sequence. Two types of classification methods, namely multi-class AdaBoost and SVM, are used for facial emotion recognition.

For appearance features, Happy *et al.* [6] proposed the use of a local binary pattern (LBP) histogram extracted from the global face region as feature vectors to be utilized for classifying facial emotions through principal component analysis (PCA). Nonetheless, this method experienced accuracy degradation due to its inability to reflect local variations of facial components in feature vectors. The fact that different face regions have varying importance levels in facial expression analysis (e.g., eyes and mouth carry richer visual information for FER as compared to forehead and cheek) is leveraged by Benitez-Quiroz *et al.* [7], who divided the human face region into several domain-specific local regions via incremental search technique and extracted the region-specific appearance features. Such an approach can effectively reduce the dimensions of resulting feature vectors while achieving significant improvement in FER accuracy. In certain cases, like [8] and [9], the hybrid of geometric and appearance features is found to complement each other's weaknesses and produce better recognition results than their standalone counterpart. Commonly, conventional approaches are known to be less computationally intensive as compared to deep learning-based approaches and therefore, preferable for applications that need real-time performance in the embedded system environment. However, since both the design of the feature extraction process and classifiers are manually and separately done by the programmer/engineer, it is not possible to jointly optimize the two processes for performance improvement. Furthermore, the majority of these handcrafted features are FER application-specific and have relatively poor generalizability in environmental conditions with large variations in lighting, subjects' ethnicity, and image resolution.

The advancement of deep learning (DL) algorithms such as convolutional neural networks (CNN) and recurrent neural networks (RNN) over the years has brought great breakthroughs to various computer vision applications, including object classification, face detection, scene understanding, and FER. Unlike conventional approaches which exhibit the deficiency of feature confusion between facial expression and facial identity due to handcrafted features [9], the feature-learning architecture of the deep neural network is able to overcome such confusion and therefore achieve better generalizability in FER. The use of CNN in FER can enable end-to-end learning directly from the input facial images, thus greatly reducing or totally removing the dependency on physics-based models and other pre-processing methods that are often found in conventional FER approaches. The capability and potential of neural networks in detecting emotions have been demonstrated by Breuer and Kimmel [10] who adopted CNN visualization techniques to understand a trained model with different FER datasets. In [11], two different CNNs namely the deep temporal appearance network and deep temporal geometric network are used to extract appearance features and geometric features from the image sequences and facial landmark points respectively. A joint fine-tuning method is introduced to integrate the two networks for improving the FER performance. Other than that, unified deep neural networks called deep regions and multi-label learning (DRML) networks were introduced by Zhao *et al.* [12] for multi-label AU detection with a newly proposed region layer. The purpose of the region layer is to capture local appearance changes of different facial regions which have been proven to provide unique cues for recognizing AUs and holistic expressions [13]. The resulting end-to-end trainable network is able to automatically learn feature representations that are invariant to changes inherent within a local facial region.

Both the aforementioned works either employ existing CNNs with modifications or formulate new CNN architecture and use them directly for FER. However, a major drawback of such approaches is that CNN can only be used for spatial feature extraction in input data but is unable to reflect the temporal relations between facial components in frame sequences. More recent works have proposed the use of a hybrid approach that combines CNN with long short-term memory (LSTM) to jointly extract spatial features of individual frames and temporal features of frame sequences. LSTM is a special type of RNN with a chain-like structure that uses short-term memory to overcome the long-term dependency problem in RNN. Kim *et al.* [14] proposed a spatio-temporal feature representation learning method that utilizes representative expression states (i.e., onset, apex and offset of facial expressions) to achieve FER that is robust against expression intensity variations. The spatial facial feature representations of the expression-state frames are learned using a CNN

and the resulting spatial features are forwarded to an LSTM to extract the temporal characteristics of facial expressions. The proposed method is evaluated with extensive experiments on both deliberate expression and micro-expression datasets and demonstrated better FER rates than previous state-of-the-art deep learning-based methods. Hasani and Mahoor [15] explored the spatial and temporal relations within facial images in video frames with a 3D Inception-ResNet architecture concatenated with an LSTM unit. To emphasize the importance of key facial components that contribute more significantly to facial expressions, facial landmark points are included as inputs to the network. A multi-angle optimal pattern-based deep learning method is recently proposed [16] to tackle the challenge of sudden illumination changes and find the proper alignment of feature sets with multi-angle-based optimal configurations. Extended boundary background subtraction is carried out to isolate foreground regions from facial images which minimizes illumination and pose variation. The texture patterns and relevant facial features are then extracted, and the facial expressions are predicted with an LSTM-CNN.

Many of the aforementioned hybrid approaches have achieved significant performance improvement as compared to ordinary CNN-based approaches by exploiting the expression dynamics (i.e., spatio-temporal features) in facial expressions. On the other hand, these deep learning-based FER approaches require a large amount of training image data in order to extract optimal facial emotion features. More importantly, deep learning-based approaches inevitably consume much higher computational power when compared to conventional approaches regardless of the training or testing process. As such, it is a research challenge to lower the computational cost of such approaches, especially during inferencing time meanwhile preserving their superiority in recognition accuracy.

## 2. THE PROPOSED METHOD

The dynamic pattern of facial expressions is comprised of three phases, namely onset (i.e., the beginning of an expression), peak (a.k.a. peak, the maximum intensity of an expression), and offset (i.e., the moment when an expression disappears). On many occasions, the changes in facial expressions from the onset phase to the offset phase happen just in a short period of time, thus making the process of real-time facial expression recognition a very challenging task [17]. With the rapid advancements of deep neural networks (DNN) in recent years, the use of deep learning approaches in the field of facial emotion recognition has become increasingly popular as they produce more promising results with higher accuracy and reliability. Compared to conventional approaches which use self-engineered image features to train the emotion classifier, DNN-based approaches are able to harvest richer and more discriminative image features in the human face, thus allowing the trained classifier to learn a better interpretation of different facial expressions.

In this paper, a facial emotion recognition pipeline is formulated by concatenating a pre-trained face detector with a fine-tuned DNN-based facial emotion classifier. An intermediate face alignment stage is incorporated in between the face detection and emotion classification stages to obtain a canonical alignment of the detected face before it is forwarded to the classifier. Several well-known and state-of-the-art CNN architectures are used for the training process of emotion classifiers to identify the network architecture which gives the best speed-accuracy trade-off. In addition, two different facial emotion datasets (i.e., unaligned color images versus aligned grayscale images) are prepared and trained with the best-performed network architecture to investigate the effects of colors and uniformity of facial images on the resulting recognition accuracy. Both image and video-based evaluations are carried out with isolated testing datasets to assess the recognition rate of the proposed facial emotion recognition system. Evaluation results show that the use of aligned grayscale images as a training dataset yields higher accuracy particularly in the video-based evaluation and the lightweight MobileNet [18] is identified as the network architecture which provides the best speed-accuracy trade-off. The obtained results allow future FER applications to be made more cost-effective (i.e., with lower computational cost) while still capable of providing real-time performance.

## 3. RESEARCH METHOD

### 3.1. Dataset preparation & model training

Two distinct facial image training datasets are prepared by the authors to examine the influence of image colors and facial alignment on the recognition accuracy of the resulting FER models. A total of six publicly available face emotion datasets are utilized to form the two training datasets, namely Facial Expression Recognition (FER-2013) dataset [19], EMOTIC dataset [20], Indian Movie Face Database (IMFDB) [21], KDEF and AKDEF datasets [22], [23], Extended Cohn-Kanade Dataset (CK+) [24] and BAUM-2 dataset [25]. Each of the six datasets is annotated with the six basic facial expressions. In order to mitigate the dataset bias problem, the authors manually screen through all the datasets to eliminate facial images that are mislabeled and those images that show nuanced expressions (i.e., only facial images that exhibit an emotion explicitly are

preserved). To further improve the uniformity of facial images in each emotion class, images with partially visible faces (e.g., the left half of the face, and upper half of the face) are also omitted.

Due to the huge unbalance in the image number of the six emotion classes, the resulting training image datasets are only made up of three emotion classes, namely angry, happy and sad that have a relatively higher and closer number of images with respect to each other. This step is taken to avoid the possible problem of the trained model's overfitting on the emotions with more image samples that would deteriorate the recognition accuracy. The first facial image training datasets are made up of color images and the images are used directly without any other pre-processing. The second dataset is derived from the first dataset by performing grayscale conversion and facial alignment. The operation flow of the facial alignment process is described in Figure 1. Facial landmark detection is first carried out to obtain the landmark points of facial components. The facial landmark detector is implemented based on [26]. Next, the centroid of both eyes as well as the angle between the two centroids are computed. The rotation matrix is then created based on the angle in the previous step after calculating the midpoint between the eyes. The midpoint is situated at the top of the nose, and it will be the reference point at which the face is rotated around. Lastly, an affine transformation is applied to align the face such that the eyes lie on a horizontal line (i.e., the centroid of eyes lie on the same y-coordinates).

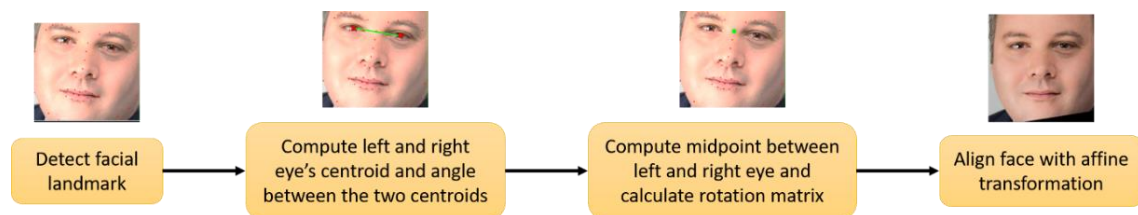


Figure 1. Facial alignment process flow

Facial images that failed the facial alignment process (i.e., facial landmarks cannot be detected) are omitted, hence the total number of images per emotion class in the resulting second dataset appeared to be smaller compared to the first dataset. Some image samples for each of the three emotion classes in both training image datasets are shown in Figure 2. The number of image samples for each class in the two datasets is stated in Table 1.



Figure 2. Image samples of training image datasets

Table 1. Number of image samples in training image datasets

	Number of Image Samples		
	Angry	Happy	Sad
Training Dataset 1: Color & Unaligned Face	2218	2340	1950
Training Dataset 2: Grayscale & Aligned Face	1359	1496	1240

The training process is carried out using several different state-of-the-art CNN models by applying transfer learning with Tensorflow deep learning framework version 1.14. A machine running Ubuntu 16.04 OS and equipped with GTX1080 GPU with 8 GB DDR5 RAM, 8th-gen i7-8700K CPU with 6 cores of up to 4.6 GHz is used for the training process. The CNN models used include Inception\_v3 [27], NASNET-large [28], ResNet-v2\_152 [29], MobileNet\_v1 [18], and MobileNet\_v2 [30]. Each model is fine-tuned and trained for 20,000 steps with an initial learning rate of 0.01 and batch size of 100 images. For validation purposes, 20% of each training dataset is used. To identify the CNN model that can provide optimal speed-accuracy trade-off, preliminary testing is performed on the fine-tuned classification models trained with the larger and more challenging training dataset 1 (i.e., due to varying color intensity and random face alignment). A relatively small, isolated testing dataset with 71, 89, and 90 image samples for the angry, happy, and sad emotion classes respectively are prepared for the preliminary testing. Evaluation results are shown in Table 2. Among the 5 models, NASNet-large gives the highest overall accuracy of 88%. The model with the worst performance is ResNet-v2\_152 with an overall accuracy of 83%. The accuracy of all models is consistently above 80%.

Table 2. Classification accuracy of trained model on preliminary testing dataset

CNN Model	Accuracy (%)			Overall
	Angry	Happy	Sad	
Inception-v3	87.32	84.27	84.44	85.20
NASNet-large	87.32	87.64	91.11	88.80
ResNet-v2_152	81.69	85.39	83.33	83.60
MobileNet-v1_100_224	85.92	89.89	85.56	87.20
MobileNet-v2_100_224	84.51	89.89	82.22	85.60

The five trained models are also evaluated in terms of their computation cost. The computation time for each model is measured using the same machine used for the training process. Table 3 shows the computation time of each model to classify a single query image. It can be observed that the MobileNet\_v1 gives the best speed-accuracy trade-off with the shortest computation time and just a 1.6% decrement in accuracy as compared to NASNet-large.

Table 3. Computation time of trained models

CNN Model	Computation time per query image (ms)
Inception-v3	14.65
NASNet-large	46.59
ResNet-v2_152	19.65
MobileNet-v1_100_224	2.4
MobileNet-v2_100_224	3

### 3.2. Facial emotion recognition pipeline

The operation flow of the proposed facial emotion recognition pipeline is depicted in Figure 3. The process starts off by reading frames from the input video stream. Face detection is then performed on each frame with the use of a pre-trained deep-learning face detector from the “dnn” module shipped with OpenCV. The face detection model is basically a ResNet10-SSD caffemodel that has been widely used to achieve fast and robust face detection. Each detected face region is then cropped out based on their respective bounding box specified as a pair of (x-y) coordinates. Next, facial landmarks detection is performed followed by facial alignment to obtain a face image patch that is horizontally aligned. The detailed description of the facial alignment process is the same as the one used in the dataset preparation process (i.e., training dataset 2). Finally, the aligned face image patches are forwarded as input to the trained CNN-based FER classifier to estimate the face emotion class (i.e., angry, happy, or sad).

To prevent prediction flickering, which is common in video classification tasks, a rolling prediction averaging mechanism is adopted to keep a list of the last K predictions (i.e., the prediction results of the last K frames) output from the FER classifier. The final stabilized emotion label for each detected face region is decided based on the emotion class that shows the highest confidence level. The authors found that a K value

of 60 is optimal for the task to produce stable results with minimum detection latency. To visualize the FER results, the predicted emotion label is overlaid on each frame at the top left corner of each detected face region highlighted by a bounding box and consequently converted into video output with a video writer. The entire FER pipeline is coded with Python 3 programming language with the help of an image processing library (i.e., OpenCV 3), deep learning library (i.e., Tensorflow v1.1.4), and utility module (i.e., VideoStream helper function in imutils package).

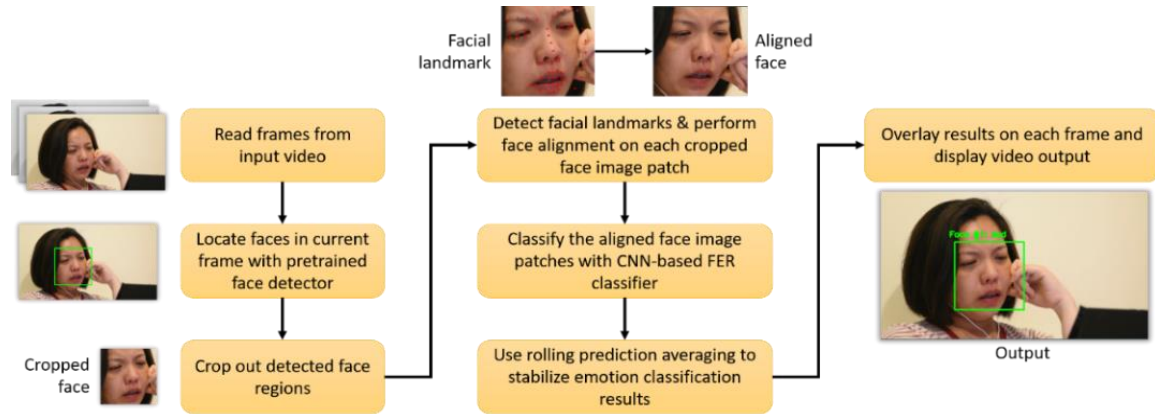


Figure 3. Overview of proposed facial emotion recognition pipeline

#### 4. RESULTS AND DISCUSSION

Two different types of testing datasets are prepared for assessing the recognition accuracy of the trained FER model, namely a testing image dataset comprised of 553 facial images and a testing video dataset comprised of 162 video sequences. All Image samples in the testing image dataset are isolated and do not overlap with the training image datasets. Of the video samples in the testing video dataset, all of them are collected online from free stock video websites such as Videezy.com, Pixabay.com, Pexels.com, and Videvo.net. The length of video samples ranges from 4 seconds up to 30 seconds. Each video contains only one subject and is labeled with just one main facial emotion (i.e., either angry, happy, or sad). The exact number of testing samples for each emotion class of the two testing datasets can be found in Table 4.

Table 4. Number of testing samples for each emotion class

Testing Dataset	Number of Samples			
	Angry	Happy	Sad	Total
Testing Image Dataset	171	246	136	553
Testing Video Dataset	22	90	50	162

To examine the performance difference of the FER model when trained with the two different training image datasets (i.e., color-unaligned facial image dataset versus grayscale-aligned facial image dataset), the MobileNet\_v1 (i.e., the CNN that is identified to provide optimal speed-accuracy trade-off) is retrained on the two training datasets with their two hyper-parameters being adjusted iteratively to determine the set of a parameter value that can generate the best performing model. The two hyper-parameters are: i) width multiplier (WM) with a value set of (0.25, 0.50, 0.75, 1.00) and ii) resolution multiplier (RM) with a value set of (128, 160, 192, and 224). The function of WM is to uniformly prune each network layer by reducing the number of input and output channels. RM is applied implicitly to a network to lower the input resolution by adjusting the image resolution. The best-performing model on the two respective training datasets is identified by observing the combination of hyper-parameters values that gives the highest final test accuracy generated at the end of the training process. Details of the two models are shown in Table 5.

The evaluation results of the two best-performing models on both image and video testing datasets are depicted in Tables 6 and 7 respectively. By observing the results, it can be found that the angry class is the most challenging emotion for FER as its accuracy is consistently lower than the other two emotion classes regardless of the type of testing datasets. For the testing image dataset, MobileNet\_v1\_1.00\_224 significantly outperforms MobileNet\_v1\_0.75\_160 for both happy and sad classes while the latter can achieve better

recognition rate for the most challenging angry emotion class. The overall accuracy of both trained models on the testing image dataset is comparable to each other with the MobileNet\_v1\_0.75\_160 demonstrating more balanced performance (i.e., the difference between the accuracy level across three classes is smaller).

Table 5. Details of best performing model

Training Image Dataset	Best Performing Model with Corresponding Hyper-parameter Value
Training Image Dataset 1 (Color-unaligned facial images)	MobileNet_v1_1.00_224 WM: 1.00, RM: 224, 569M Multi-Adds Computation, 4.24M parameters
Training Image Dataset 2 (Grayscale-aligned facial images)	MobileNet_v1_0.75_160 WM: 0.75, RM: 160, 162M Multi-Adds Computation, 2.59M parameters

Table 6. Evaluation results on testing image datasets

CNN Models	Accuracy (%)			
	Angry	Happy	Sad	Overall
MobileNet_v1_1.00_224 (Training Image Dataset 1)	35.67	91.06	73.53	69.62
MobileNet_v1_0.75_160 (Training Image Dataset 2)	40.35	86.18	64.70	66.72

Table 7. Evaluation results on testing video datasets

CNN Models	Accuracy (%)			
	Angry	Happy	Sad	Overall
MobileNet_v1_1.00_224 (Training Image Dataset 1)	72.73	81.11	92.00	83.33
MobileNet_v1_0.75_160 (Training Image Dataset 2)	77.30	84.40	94.00	86.42

As for the testing video dataset, the MobileNet\_v1\_0.75\_160 which is trained on the grayscale-aligned facial image dataset consistently outperforms the MobileNet\_v1\_1.00\_224 for all three emotion classes. This finding implies that the MobileNet\_v1\_0.75\_160 can provide better generalizability in video-based FER tasks when the facial expression is transiting from the onset phase to the apex phase and before the offset phase. The fact that MobileNet\_v1\_0.75\_160 which is trained with lower-resolution facial images (hence lesser features) can perform better suggests that human generally recognizes facial expression in a holistic way without relying much on facial features [31]. As such, people who have impaired high-frequency vision will still have no difficulty distinguishing different facial expressions as they can just rely on low-frequency visual information. In addition, color information is shown to be less useful in the FER task compared to the recognition of natural images as the MobileNet\_v1\_0.75\_160 trained with grayscale facial images achieves higher overall recognition accuracy. The presence of color information could possibly introduce unwanted distributional shifts in feature space that can negatively impact the recognition rates due to the unwanted features focusing on different skin colors or facial images taken under varying lighting conditions. To sum up, training a CNN-based FER model with grayscale-aligned facial images is preferable as this offers better recognition rates while reducing the detection latency (i.e., due to the smaller resolution of the input image and CNN model).

## 5. CONCLUSION

This paper presents the implementation of a deep learning-based facial emotion recognition pipeline to predict the emotion of facial images found in video sequences. Experimental results show that the lightweight MobileNet\_v1 can provide the best speed-accuracy trade-off for the FER task. In addition, using grayscale and properly aligned facial images as a training dataset was found to produce a better trained FER model compared to the use of color images with random face rotation and translation. In future research, gesture recognition can be combined together with the proposed FER pipeline to collectively estimate human emotion in a more stable and reliable manner.




## REFERENCES

- [1] K. Kaulard, D. W. Cunningham, H. H. Bühlhoff, and C. Wallraven, "The MPI facial expression database — a validated database of emotional and conversational facial expressions," *PLoS ONE*, vol. 7, no. 3, Mar. 2012, doi: 10.1371/journal.pone.0032321.
- [2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971, doi: 10.1037/h0030377.
- [3] P. Ekman and E. L. Rosenberg, *What the face reveals basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press, 2005, doi: 10.1093/acprof:oso/9780195179644.001.0001.
- [4] L. Zhang, D. Tjondronegoro, and V. Chandran, "Representation of facial expression categories in continuous arousal–valence space: Feature and correlation," *Image and Vision Computing*, vol. 32, no. 12, pp. 1067–1079, Dec. 2014, doi: 10.1016/j.imavis.2014.09.005.
- [5] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, Jun. 2013, doi: 10.3390/s130607714.
- [6] S. L. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," in *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, Dec. 2012, pp. 1–5, doi: 10.1109/IHCI.2012.6481802.

- [7] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 5562–5570, doi: 10.1109/CVPR.2016.600.
- [8] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7803–7821, Mar. 2017, doi: 10.1007/s11042-016-3418-y.
- [9] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Computer Vision -- ACCV 2014*, 2015, pp. 143–157, doi: 10.1007/978-3-319-16817-3\_10.
- [10] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *Prepr. arXiv1705.01842*, May 2017.
- [11] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2983–2991, doi: 10.1109/ICCV.2015.341.
- [12] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action Unit detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 3391–3399, doi: 10.1109/CVPR.2016.369.
- [13] S. Shojailangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh, "Robust representation and recognition of facial emotions using extreme sparse learning," *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2140–2152, Jul. 2015, doi: 10.1109/TIP.2015.2416634.
- [14] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, Apr. 2019, doi: 10.1109/TAFFC.2017.2695999.
- [15] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 30–40.
- [16] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognition Letters*, vol. 139, pp. 157–165, Nov. 2020, doi: 10.1016/j.patrec.2017.06.025.
- [17] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing lower face action units for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 484–490, doi: 10.1109/AFGR.2000.840678.
- [18] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *Prepr. arXiv1704.04861*, Apr. 2017.
- [19] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, 2013, pp. 117–124, doi: 10.1007/978-3-642-42051-1\_16.
- [20] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1960–1968, doi: 10.1109/CVPR.2017.212.
- [21] S. Setty *et al.*, "Indian movie face database: A benchmark for face recognition under wide variations," in *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, Dec. 2013, pp. 1–5, doi: 10.1109/NCVPRIPG.2013.6776225.
- [22] D. E. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces - KDEF," Psychology section, Karolinska Institutet, 1998.
- [23] D. F. Lundqvist and J. E. Litton, "The averaged Karolinska directed emotional faces - AKDEF," Psychology section, Karolinska Institutet, 1998.
- [24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.
- [25] C. Eroglu Erdem, C. Turan, and Z. Aydin, "BAUM-2: a multilingual audio-visual affective face database," *Multimedia Tools and Applications*, vol. 74, no. 18, pp. 7429–7459, Sep. 2015, doi: 10.1007/s11042-014-1986-2.
- [26] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1867–1874, doi: 10.1109/CVPR.2014.241.
- [27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [28] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8697–8710, doi: 10.1109/CVPR.2018.00907.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016, pp. 630–645, doi: 10.1007/978-3-319-46493-0\_38.
- [30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [31] A. O. Omigbodun and G. W. Cottrell, "Is facial expression processing holistic?," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 2013, pp. 3187–3192.




## BIOGRAPHIES OF AUTHORS






**Ng Chin Kit**    received his B.Eng. (Hons) and his Master of Engineering Science from Multimedia University, Malaysia. He is currently a software engineer at META Research Sdn. Bhd., a multinational company focusing on retail technology development. His career and research interests include deep learning for computer vision applications, machine learning pipeline architecture design, native mobile app development, and edge computing. He can be contacted at [chinkitng@yahoo.com](mailto:chinkitng@yahoo.com).








**Chee-Pun Ooi**    obtained his Ph.D. in Electrical Engineering from University of Malaya, Malaysia, in 2010. Dr. Ooi's current position is a senior lecturer at Faculty of Engineering, Multimedia University. He is a chartered engineer registered with Engineering Council (the British Regulatory Body for Engineers) and a member of IET. His research areas are in FPGA-based embedded systems and embedded systems. He can be contacted at email: cpooi@mmu.edu.my.






**Wooi Haw Tan**    received his M.Sc. in Electronics from Queen's University of Belfast, UK and a Ph.D. in Engineering from Multimedia University. He is currently a senior lecturer at Multimedia University. Dr. Tan's areas of expertise include image processing, embedded system design, Internet of things (IoT), machine learning, and deep learning. He has participated in various government and privately funded projects since he started his career with the University. His works have been published in numerous international journals and conferences. Besides, he has also co-authored two textbooks on microcontroller systems. Dr. Tan has been actively involved in designing and developing hardware prototypes for IoT based embedded systems and software systems for machine learning and deep learning. He can be contacted at email: wktan@mmu.edu.my.



**Yi Fei Tan**    obtained her B.Sc. (Hons), M.Sc., and Ph.D. from University of Malaya (UM). She is currently a senior lecturer at the Faculty of Engineering, Multimedia University (MMU), Cyberjaya, Malaysia. Her research interests include machine learning, deep learning, image processing, big data analytics, and queueing theory. She can be contacted at email: yftan@mmu.edu.my.



**Soon Nyeon Cheong**    received his B.Eng. (Hons) and his Master of Engineering Science from Multimedia University, Malaysia. He is currently a senior lecturer at Faculty of Engineering, Multimedia University, Malaysia. He is a reviewer for a number of international journals and conferences. He has received grants from Telekom Malaysia, Penang ICT, MOSTI, and MOHE for his research works. Soon Nyeon Cheong has published research papers in the form of books, book chapters, peer-reviewed journals, and international conference proceedings. His teaching and research interests include web engineering, natural user interface, smart home, gerontechnology, educational technology, and interactive multimedia content. He can be contacted at email: sncheong@mmu.edu.my.