A Development Framework for Automated Facial Expression Recognition Systems

Bacha Rehman[1], Wee Hong Ong[2] , Trung Dung Ngo [3]
[1] Department of Computer Science, Namal Institute, Mianwali, Pakistan
[2] Faculty of Science, Universiti Brunei Darussalam, Brunei Darussalam
[3] The More-Than-One Robotics Laboratory, University of Prince Edward Island, Canada
bacha.rehman@namal.edu.pk[1], weehong.ong@ubd.edu.bn[2], dungnt@ieee.org[3]

Abstract. Automated facial expression recognition (AFER) has become an important research area with several computer vision (CV) applications. A robust AFER system requires sufficient good quality training and testing data for development and evaluation of a robust AFER model. There exist a number of AFER datasets and an increasing number of research works in AFER. However, research works in AFER have not matured to a stage that there are openly available platforms or toolsets to implement the pipeline of AFER system development. New comers to the field are faced with various challenges such as 1) images in the datasets are messy or with low resolutions; 2) the data are not organized into separate training and testing data for fair evaluation; 3) majority of the datasets are very small leading to insufficiency for training a model; 4) some datasets do not provide important facial features, 5) it is unclear which dataset to start with, and 6) no development framework and methodologies to systematically implement and test new models. In this paper, we present a framework with complete source code and algorithms to: 1) detect faces and crop face images in a given dataset for AFER; 2) extract facial landmark features from the face images and store as landmark images; 3) split the dataset into training and testing sets and stored into two CSV files consisting filename, emotion, landmarks, and features vectors for each image in its respective set; 4) train and evaluate the features vectors of the dataset using a deep neural network (DNN) model as the baseline; 5) train and evaluate a baseline convolutional neural network (CNN) on the face cropped images; 6) demonstrate the trained model on live videos and images. This study also outlines the necessary steps involved in developing an AFER system. The framework can help researchers to use a dataset to develop AFER systems and further improve the framework and benchmark the results.

Keywords: Facial Keypoints; Facial Expressions; Emotions Recognition; Deep Neural Network; Convolutional Neural Network; Facial Expressions Dataset

1 Introduction

Facial expressions are vital for affect signaling systems. It is the natural source of cues regarding a person's emotional state. Significant research has been recently done in the area of AFER. Several datasets have been created as well as different methods are applied to develop state of the art AFER systems [1]. Different machine learning models have been developed for AFER systems including convolutional neural network (CNN), deep neural network (DNN), support vector machine (SVM), local binary patterns (LBP), hybrid CNN + recurrent neural network (RNN), hybrid principal component analysis (PCA) + Latent Dirichlet allocation (LDA), and support

vector regression (SVR) based action units (AU) intensity [1].

In this paper, we have utilized CNN and DNN based deep learning models as part of the framework provided in this paper. It is because the CNN models are one of the most popular deep learning models used for images and face related research. However, sometimes the CNN models fail to be adapted for real-world applications due to its robustness mainly caused by insufficient training data, simplified architecture of a particular CNN and the limited variance and diversity in the training data. Therefore, it is essential that researchers should know about alternative deep learning model, i.e. DNN in this paper. A DNN based baseline model is also provided as part of the development environment in this work so that researchers should adapt and use features based deep learning methods.

Facial expression datasets are essential to train machine learning algorithms for an AFER system. The accuracy of the training outcome is highly affected by the size and diversity in a dataset. There are many relevant facial images datasets [2], however, a few of them are conveniently useful to develop an AFER system. Finding and selecting suitable dataset is a challenging task especially for researchers new in the AFER domain. There are several challenges while selecting a database to train and test a model for AFER system:

- Some datasets are messy and have inaccurate image labels.
- Some datasets do not provide the facial landmark features useful for the development of AFER models.
- Several datasets consist of only a few hundred images and is not suitable to train a model especially using deep learning.
- Many datasets consist of only images for training without pre-defined sets of training and testing data to fairly evaluate and compare the trained model.
- The datasets usually do not provide sufficient details and tools to develop AFER systems with the datasets.

These limitations lead us to provide a development framework including DNN and CNN models, along with the suggestion of a posed expressions dataset that can also be used as spontaneous expressions dataset. This framework and the suggested dataset can suffice to allow AFER researchers to quickly develop a baseline AFER system and for further analysis and development. This paper explains the method adapted to create an organized AFER dataset from the extended Cohn-Kanade (CK+) dataset. The resulting dataset has eight facial expressions including 'Neutral', 'Angry', 'Disgust', 'Fear', 'Happy', 'Sad', 'Surprise', and 'Contempt'. The methodology behind the proposed framework has been explained in Fig. 2, while the proposed deep learning models have been elaborated in Section 5. The proposed framework includes:

1. Source code for image preprocessing, generate 68 facial landmarks, calculate distance vectors among every facial point and to save cropped images, landmark images and generate CSV file.
2. Source code to build, train, evaluate and save the baseline DNN and CNN models on the data provided in the CSV files and generate graphs for loss and confusion matrix.
3. Source code to apply the trained models on unseen images and videos including live video input.

The key contributions of this study are as follows:

- Provide a comprehensive development framework and source code for

researchers to understand, implement and evaluate AFER systems.

- Describe the development pipeline including dataset preparation based on a suggested dataset.
- Provide details of how to expand the dataset and use it for both posed and spontaneous purposes.

2 Related Work

Several face related datasets are available for researchers to conduct experiments on AFER [1], face recognition, and facial attribute analysis [3]. However, each of them was created for different purposes and they lack the details and tools to ease its usage. It is a difficult task for a new researcher group in the field to find a suitable dataset to develop AFER models. The Japanese Female Facial Expression (JAFFE) [4], Maja-Michel-Ioannis (MMI) [5], facial expression recognition (FER2013) [6], the Karolinska directed emotional faces (KDEF) [7], and the Yale [8] are well known facial expressions databases available for research purposes. The main limitations of these datasets are that they did not provide any framework or tool to use effectively use the dataset and the datasets are not easily extendable. Furthermore, the datasets are either messy or have very limited number of images which is not suitable to train state of the art deep learning models, i.e. CNN Models.

A comprehensive dataset was first provided by Cohn-Kanade (CK) [9]. This dataset was created in the form of images sequences. Each images sequence was labeled using facial action coding system FACS [10] for the corresponding facial expression. The main drawback of this dataset was that the images sequences were unverified against the actual facial expressions they comprehended. Another drawback was the lack of intensity labels, which leaded this dataset to be extended to CK+ [11]. More samples of spontaneous expressions were added to CK+. The total number of images sequences in CK+ is 593, in which 327 images sequences have discrete facial expression labels. Images resolution of CK+ is 640 × 490. Table-1 summarizes the description and limitations of all these datasets.

Table 1: Overview of available AFER Datasets (Image-Based)

Dataset	**Images / videos**	**Limitations**
FER 2013 [6]	28709 – training 3589 - test /valid each	The dataset is for competition hence very messy Images resolution is low Some emotion images reveal false facial expressions
CK [9]	2000 image sequences	Only image sequences Images sequences were unverified against the actual facial expressions
CK+ [11]	593 image sequences	Only image sequences 327 out of 593 image sequences are annotated Provided for spontaneous emotion analysis
KDEF [7]	490 images	Limited and small dataset Cropped images provided
JAFFE [4]	213 images	Limited and small dataset Only 213 images provided
MMI [5]	1500 images and image sequences	Included non-frontal faces Facial expressions added later,
Yale [8]	165 images	Small dataset not suitable for training a model

In development framework provided in this paper, we have used MTCNN [12] for face detection within the image frames because of its superior performance [13] [14] in comparison to Haar based algorithm [15] for face detection. For the facial landmark [3] extraction, we have used the DLIB library [16]. The experiments data, environments and settings were kept consistent for all the experiments in this work for uniform analysis and fair evaluation. The following deep learning models along with its given name and references were used to compare accuracy results with the proposed deep learning models in this work.

1. CNN-1 (Mini Xception) [17]
2. CNN-2 (Deep vs Shallow) [18]
3. CNN-3 (Deep CNN) [19]

3 Dataset Preparation

The CK+ dataset [11] was first divided into separate training and testing sets at a ratio of around 2:1. We derived each of the facial expressions from the available 593 sequences of the CK+ dataset. There are only 327 sequences which were annotated by the creator of CK+ dataset. We manually annotated the remaining 266 images sequences by thoroughly examining the nature of expression. There are 8 facial expressions including 'Neutral', 'Angry', 'Disgust', 'Fear', 'Happy', 'Sad', 'Surprise', and 'Contempt'. The images for each expression are organized into their respective folder.

In the CK+ dataset image sequences, the intensity of facial expression gradually increases frame-wise from Neutral towards the annotated facial expression. Keeping this in mind, the Training and Testing set have been created by taking images from the image-sequence. For instance, 1st and 2nd images in each sequence were taken as Neutral facial expression for the Training set, because each image sequence starts with a neutral facial expression. As the last two images in the sequence have the extreme visibility for the annotated facial expression, i.e. peak expression, therefore, the last two images were considered as the samples of the corresponding expression for the training set. For the testing set, the 3rd image in each sequence was taken as Neutral expression. For the other expressions, the visibility was determined from visual inspection. The search for the visibility initiated from the middle image in the images sequence until the 3rd last image of the sequence. The image with the most visible expression was taken as a sample for the test set. It was observed that the expression visibility was quite obvious for the 5th or 6th position from the last image in each sequence.

Table 2: Images details of the Dataset provided

Expression	Sub-Folder	Training Set	Testing set
Neutral	0	1174	586
Angry	1	112	56
Disgust	2	130	65
Fear	3	144	71
Happy	4	192	95
Sad	5	174	86
Surprise	6	192	96
Contempt	7	36	20

The total number of images in the resulting training set was 2154 while the testing set consisted of 1075 images as shown in Table 2. The resulting dataset can be extendable and more images from other datasets can be added into any facial expression to increase the number of training or testing samples. The workflow created an improved dataset with original facial expression images (640 x 490), cropped images (200 x 200), and facial landmark images (200 x 200 with 68 facial points). Each image was cropped around the face area detected using Multi-task Convolutional Neural Networks (MTCNN) [12]. The facial landmark [3] based images consist of 68 facial points obtained using Dlib library [16]. A CSV file was generated for the training and testing sets each. There were four fields in each CSV file, i.e. filename, emotion, landmarks, and landmarks distance vectors.

4 Dataset Pre-processing and Features Extraction

Once the images have been selected and placed into their respected folders, the images were processed and features were extracted so that the dataset can be conveniently utilized for any DNN or CNN models training. The process consists of mainly four tasks, i.e. face detection, face area cropping, facial landmarks distance vectors calculation, and save the cropped images, landmarked images and features vectors.

Once the face is detected through MTCNN, the next step is to calculate the 68 facial landmarks. Dlib library [16] has been used for calculating the facial landmarks. The face box returned by MTCNN is in rectangular shape. However the Dlib library requires the face box in a square shape. The forehead area of the face does not contribute much in the facial expression and determination of the facial landmarks. Therefore, the forehead area in the face box was discarded. The equations (1) to (3) were used to discard the forehead area of the rectangular face box. The delta ($\boldsymbol{\Delta h}$) is the number of row pixels to be cropped from the original face rectangle; h_o and w_o are the original height and width of the face rectangles; h_c is the cropped rectangle height; and x_o and y_o are the coordinates of the upper left corner of the original face rectangle. The region of interest (ROI) extracted from the original image i is a square face box suitable for the facial landmarks extraction.

$$\boldsymbol{\Delta h} = \boldsymbol{h_o} - \boldsymbol{w_o} \quad ..(1)$$

$$\boldsymbol{h_c} = \boldsymbol{h_o} - \boldsymbol{\Delta h} \quad ..(2)$$

$$\boldsymbol{ROI} = \boldsymbol{i}\,[\boldsymbol{y_o} + \boldsymbol{h_c}: \boldsymbol{y_o} + \boldsymbol{h_o}\,, \boldsymbol{x_o}: \boldsymbol{x_o} + \boldsymbol{w_o}] \quad ..(3)$$

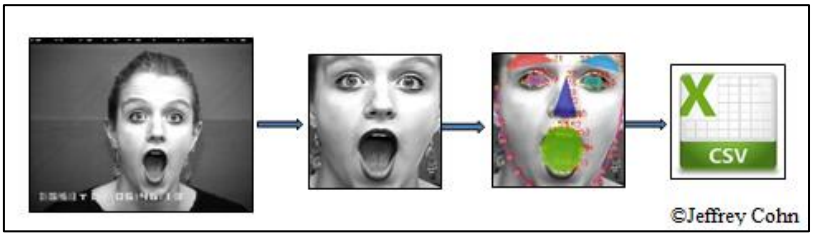

Figure 1: Pre-Processing and feature extraction for single image.

The Euclidean distance has been used to calculate the distance between all pairs of facial landmark points. Figure 1 depicts the pre-processing and feature extraction activities for a single image as described above. Figure 2 describes the complete process to create the organized dataset that has been implemented in the proposed framework.

Dataset preparation and data pre-processing process

Data: Images from a public dataset
Result: Organized dataset with folders and *CSV* files
Create folders *Train_Original*, *Test_Original*, *Train_Crop*, *Test_Crop*, *Train_LM*, *Test_LM*.
Create sub-folders *0* to *7* (eight expressions) in each of the above folders.
Split the original dataset into train and testing sets by selecting the images from image sequences, as described in Section 3, into their respective sub-folders (expressions) in *Train_Original* and *Test_Original*.
Open two new *CSV* files, *Train.csv* and *Test.csv*.
for each folder ***f*** in "*Train_Original*, *Test_Original*" **do**
  for each image ***i*** in each sub-folder in ***f*** **do**
    1. Detect the face
    2. **if** face is detected **then**
      2.1 Use Equations (1) to (3) to obtain the face box
      2.2 Crop the face box
      2.3 Resize the face box to 200 x 200
      2.4 Save the cropped image in the corresponding sub-folder in *Train_Crop* or *Test_Crop* depending on ***f***
      2.5 Extract facial landmarks
      2.6 Save landmarked images in the corresponding sub-folder in *Train_LM* or *Test_LM* depending on ***f***
      2.7 Calculate distance among all facial landmarks
      2.8 Add a line with filename, sub-folder name as emotion, landmarks, and distance vectors in *Train.csv* or *Test.csv* depending on ***f***
    end
  end
end
Close and save both *CSV* files

Figure 2: The pipeline of dataset preparation, data processing and feature extraction.

5 AFER Baseline Models

The proposed framework includes baseline models to conveniently adapt in an AFER system as well as to benchmark AFER systems developed with the same dataset. The baseline deep leaning models are DNN and CNN based classifiers. Five deep learning AFER DNN models were developed and evaluated as listed in Table 3. Based on the accuracy results, the model number 5 in Table 3 has been selected and included in the framework. Softmax has been used as the activation function in the output layer. The structure of the proposed DNN model consists of 6424 input features, 5 x 1024 hidden layers, and an output layer with 8 neurons. Min-Max features normalization has been applied for features normalization within 0 to 1 range.

Table 3: DNN Deep Learning Model Tested

#	DNN Structure	Accuracy
1	Input, 512, 512, 512, 512, 512, Output	88.55
2	Input, 512, 512, 512, 512, 512, 512, 512, Output	88.17
3	Input, 1024, 512, 512, 512, 512, Output	88.45
4	Input, 1024, 512, 512, 512, 512, 512, 512, Output	88.36
5	Input, 1024, 1024, 1024, 1024, 1024, Output	89.58

Apart from the DNN based model, a CNN based model has also been designed and included in the framework. Figure 3, shows the architecture of the baseline CNN model. The face images are resized to 48 x 48. The proposed CNN model consists of 4 convolutional layers of size 64, 128, 512 and 512 respectively. The kernel size for each convolutional layer were taken as (3 × 3), (5 × 5), (5 × 5), and (3 × 3) respectively. Each convolutional layer was followed by batch normalization, maximum pooling of pool size = (2 × 2), and a dropout layer of size 0.2 as shown in Fig 3. After the fourth dropout layer, the features vectors were flatten using the Flatten layer. Two fully connected layers of size 512 and 1024 have been used after the Flatten layer. Each fully connected layer has been followed by a batch normalization layer and a dropout layer of size 0.25. Finally the output layer was placed after the final dropout layer.

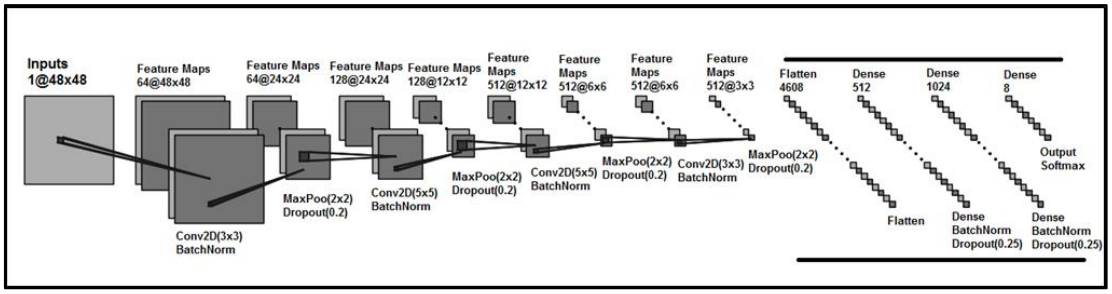

Figure 3: Architecture of the CNN baseline model.

Stochastic gradient descent (SGD) [20] was used as the model optimizer. Maximum number of epochs for the training process was set at 500. The learning rate of the SGD algorithm was set at 0.01. No stopping criteria were applied during the training process. We reduced learning rate during the training process with settings as factor = 0.5, patience = 50, and minimum learning rate = 0.0001. The batch size was set to 64.

6 Results and Discussion

Experiments were conducted using the two deep learning models described in Section 5. CPU computation has been used for both training and testing process. The results reported here are intended for benchmarking of other AFER systems developed using the suggested dataset. Figure 4 shows the loss graph and normalized confusion matrix (accuracy) for the training and testing data of the proposed DNN baseline model. Figure 5 shows the same information for the proposed CNN baseline model.

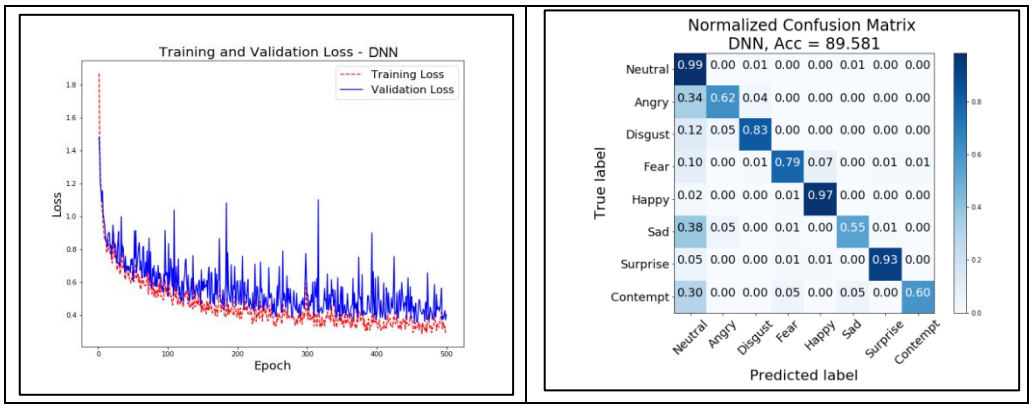

Figure 4: Loss and Normalized confusion matrix for DNN model

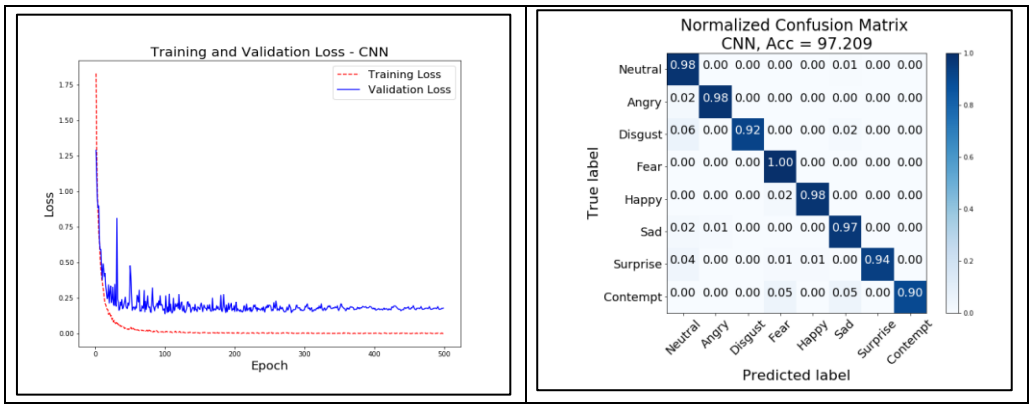

Figure 5: Loss and Normalized Confusion Matrix graph for CNN model

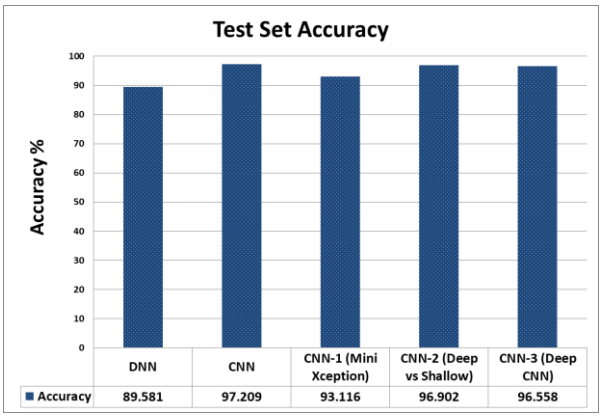

Figure 6: Classification accuracy achieved by each deep learning model over the testing set

The benchmark accuracy achieved using the CNN based model was significantly higher than the result of the DNN based model. The accuracy achieved using the CNN model was 97.21 % on the prepared testing set. However, it was observed that

the DNN based model works quite robustly on the unseen data such as the live stream from webcam. DNN models are learned from extracted features. DNN models are less biased to the nature of the images and more subjects independent. The dataset generated is sufficient to train any features based DNN model. However, the CNN model can be made more robust by adding more data or using transfer learning [21]. A demo video is provided to show the performance of the baseline DNN model on live video of unseen subject [22]. Apart from the deep learning models proposed in this work, we also applied three deep learning models presented in [17], [18] and [19] on the dataset used in this work to compare the proposed models performance with the-state-of-the-art models. Figure 6 shows the classification accuracy for all the deep learning models for comparison.

7 Conclusion

A development framework consists of DNN and CNN based models are provided as a baseline framework for implementation and development of AFER systems. The framework can be used to develop, train, and evaluate any AFER system using deep learning models to recognize the eight common facial expressions. Experiment results have shown that the feature-based DNN baseline model exhibited robust real-time performance to detect correct facial expressions for the unseen subject. The suggested dataset can be used for both posed and spontaneous facial expression analysis. Lastly, the framework have in consideration to allow easy expansion of the generated dataset with more data to be added in order to train more robust CNN based models. The proposed framework can ease the learning curve of new researchers in developing their own AFER systems. The source code files for the development framework are provided on GitLab [23]. These source codes were used to generate CSV files, develop / train / evaluate deep learning models, and finally demonstrate the trained models over unseen images and live videos.

References

1. Mehta, D., Siddiqui, M.F.H., Javaid, A.Y.: Facial emotion recognition: A survey and real-world user experiences in mixed reality. Sensors (Switzerland). 18, 1–24 (2018). https://doi.org/10.3390/s18020416.
2. http://www.face-rec.org/databases/.
3. Bulat, A., Tzimiropoulos, G.: How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017). https://doi.org/10.1109/ICCV.2017.116.
4. Lyons, M.J., Shigeru, A., Miyuki, K., Jiro, G., Budynek, J.: The Japanese female facial expression (JAFFE) database. In: Proceedings of third international conference on automatic face and gesture recognition. pp. 14–16 (1998).
5. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: WEB-BASED DATABASE FOR FACIAL EXPRESSION ANALYSIS. In: IEEE international conference multimedia and Expo. pp. 5–8 (2005). https://doi.org/10.1109/ICME.2005.1521424.
6. Goodfellow, I.J., et al.: Challenges in representation learning: A report on three machine learning contests. Neural Networks. 64, 59–63 (2015). https://doi.org/10.1016/j.neunet.2014.09.005.
7. Goeleven, E., De Raedt, R., Leyman, L., Verschuere, B.: The Karolinska directed emotional faces: A validation study. Cogn. Emot. 22, 1094–1118 (2008).

https://doi.org/10.1080/02699930701626582.
8. Georghiades, A., Belhumeur, P., Kriegman, D.: From Few To Many: Generative Models For Recognition Under Variable Pose and Illumination. IEEE Trans. Pattern Anal. Mach. Intell. 23, 643–660 (2000).
9. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG. pp. 46–53 (2000). https://doi.org/10.1109/AFGR.2000.840611.
10. Hamm, J., Kohler, C.G., Gur, R.C., Verma, R.: Automated Facial Action Coding System for Dynamic Analysis of Facial Expressions in Neuropsychiatric Disorders. J. Neurosci. Methods. 200, 237–256 (2012). https://doi.org/10.1016/j.jneumeth.2011.06.023.Automated.
11. Lucey, P., Cohn, J.F., Kanade, T., et al.: The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. pp. 94–101 (2010). https://doi.org/10.18632/aginh.101501.
12. Zhang, K., Zhang, Z., Li, Z., Member, S., Qiao, Y., Member, S.: Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. IEEE Signal Process. Lett. 23, 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342.
13. Rehman, B., Ong, W.H., Hong, A.T.C.: Using Margin-based Region of Interest Technique with Multi-Task Convolutional Neural Network and Template Matching for Robust Face Detection and Tracking System. In: Proceedings of 2nd International Conference on Imaging, Signal Processing and Communication (ICISPC). pp. 14–18 (2018).
14. Rehman, B., Ong, W.H., Tan, A.C.H., Ngo, T.D.: Face detection and tracking using hybrid margin-based ROI techniques. Vis. Comput. 1–15 (2019). https://doi.org/10.1007/s00371-019-01649-y.
15. Rehman, B., Ong, W.H., Hong, A.T.C.: Hybrid Model with Margin-Based Real-Time Face Detection and Tracking. In: The 11th Multi-disciplinary International Workshop on Artificial Intelligence (MIWAI). Lecture Notes in Computer Science. pp. 360–369. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-49397-8.
16. Baltrusaitis, T., Robinson, P., Morency, L.P.: OpenFace: An open source facial behavior analysis toolkit. In: IEEE Winter Conference on Applications of Computer Vision, WACV. pp. 1–10 (2016). https://doi.org/10.1109/WACV.2016.7477553.
17. Arriaga, O., Valdenegro-Toro, M., Plöger, P.: Real-time Convolutional Neural Networks for Emotion and Gender Classification. In: Proceedings of the 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN-2019), April 24-26, Belgium (2017).
18. Alizadeh, S., Fazel, A.: Convolutional Neural Networks for Facial Expression Recognition. arXiv Prepr. arXiv 1704.06756. (2017).
19. Yu, Z., Zhang, C.: Image based Static Facial Expression Recognition with Multiple Deep Network Learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 435–442 (2015). https://doi.org/10.1145/2818346.2830595.
20. Babenko, B., Yang, M.H., Belongie, S.: A family of online boosting algorithms. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, ICCV Workshops. pp. 1346–1353 (2009). https://doi.org/10.1109/ICCVW.2009.5457453.
21. Ng, H.-W., et al.: Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 443–449 (2015). https://doi.org/10.1145/2818346.2830593.
22. AI Lab.: Multimodal Human Intention Perception For Human Robot Interaction. [online] Ailab.space, Available at: <https://ailab.space/projects/multimodal-human-intention-perception/#visual-facial-expression-recognition> [Accessed 4 September 2020].
23. Rehman, B., Ong, W.H., Ngo, T.D.: A development framework for Automated Facial Expression Recognition Systems, GitLab project (2020). https://gitlab.com/ailab.space/bacha-afer-dev-framework.