Scientific Visualization, 2022, volume 14, number 1, pages 62 - 86, DOI: 10.26583/sv.14.1.06

Facial Emotion Recognition Through Detection of Facial Action Units and Their Intensity

Rohan Appasaheb Borgalli[1,A], Sunil Surve[2,B]

[A] Department of Electronics Engineering, Fr. Conceicao Rodrigues College of Engineering, Bandra, University of Mumbai, Mumbai, India
[B] Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Bandra, University of Mumbai, Mumbai, India

[1] ORCID: 0000-0003-4159-1938, rohanborgalli111@gmail.com
[2] ORCID: 0000-0002-2568-9911, surve@fragnel.edu.in

Abstract

Facial expression recognition (FER) is a vital process in many applications in computer vision, medical, human-computer interface, video games, AI, E-education, security, distance psychotherapy, and Counselling. In the past, only basic emotions are the focus of research but now, with advancement in technology and need along with basic emotion, reorganization of compound emotions is also getting important. But, recognizing facial expressions with accuracy is challenging. To deal with that, facial action units for detecting appropriate basic and compound facial emotion are useful.

Among state-of-the-art methods for FER systems, detection of facial action units (AUs) showed good results. Combining these AUs to detect particular basic and compound facial expressions is key and improves the accuracy of basic and compound facial expressions.

We have used standard CNN architecture VGGNet, XceptionNet, and ResNet to conduct experiments in the proposed method to explore CNN architecture that best performs for action unit recognition. We modified them slightly by changing a few final fully connected layers to detect facial action unit intensity. Using this modified architecture for CK+, MMI, DISFA, and DISFA+ database that detects action units with good accuracy, it is intern mapped to find basic and compound facial emotions. Our Proposed method achieves an overall accuracy for Action Unit detection using Xceptionnet network for MMI & DISFA are, giving promising results average F1- score is 72% and 74%, respectively. In contrast, a network for CK+ and DISFA+ has an overall F1- score is 62% for both.

Keywords: facial expression; facial action unit; convolution neural network.

1. Introduction

FACIAL expressions of human beings are indications of their emotional states and intentions and can be used as a tool to understand them [1]. In the past, Numerous attempts have been made to implement a facial expression recognition system as practically it has tremendous importance in applications in the field of machine vision and machine learning, various facial expression recognition (FER) systems have been explored to encode expression information from facial representations. Ekman and Friesen [2] defined six basic emotions based on cross-culture study as anger, disgust, fear, happiness, sadness, and surprise, which indicated that humans perceive certain basic emotions in the same way regardless of culture. Contempt was subsequently added as one of the basic emotions [3].

Recently, advanced research on FER indicates that faces display basic emotions and compound emotions, combining the six basic emotions. Compound emotion corresponds to the superposition of two basic emotions leads to different types of emotions given by Du et al. [4] as 22 emotions, including seven basic emotions, 12 compound emotions most typically

expressed by humans, and three additional emotions (appall, hate, and awe). Also, JIANZHU GUO et al. [5] Compressed problems of compound emotions has limited datasets with a limited number of categories and unbalanced data distribution labelled automatically by machine learning-based algorithms, leading to inaccuracies. They released the iCV-MEFED dataset, including 50 classes of compound emotions and labels assessed by psychologists.

Facial Action Coding System (FACS)[6] is designed which specify 46 facial muscle movement units called Action Unit (AU). It has classified its upper and lower facial action units. Upper facial action units indicate the region near the eyes and nose, whereas lower ones cover the cheek, lips, and chin.

The human facial expression consists of basic and compound emotions that cannot be simply described as combining some facial regions, especially when there are compound expressions and micro-expressions. Multiple regions of the face will individually or combined appear, which gives rise to particular facial emotion, so it is difficult for classifiers trained based on discriminant methods to describe them. Any facial expression results from the involvement of group of facial muscles and as a combination of several AUs. Compared with the information judgment method, FACS can quantitatively measure and evaluate facial movement, making it more objective and comprehensive.

Compound emotions are defined as different combinations of basic emotion and in this combination, how AU plays the role is crucial. Sometimes the particular AU's is present in emotion, but its intensity is very low. Such expressions are called micro-expressions, and they are complicated to recognize as they are very subtle in nature. All this leads to a challenging situation in any Facial Expression recognition system based on static images.

Though much research has been done, recognizing facial expressions with a high accuracy remains to be challenging due to the complexity and varieties of facial expressions. We tried to deal with this problem by proposing different architectures suitable for specific datasets to detect Facial action units accurately. With the help of mapping Action units with emotion given in [5], basic and compound emotion detection is identified.

2. Related work

To model, a Facial Expression Recognition system based on basic emotions is not sufficient to represent the complication of human facial expressions precisely. To deal with this, mainly two emotion description models are designed. The first one is the Facial Action Coding System(FACS)[6], which is a human-observer-based system to describe subtle changes in facial features as a continuous model using affect dimensions [7]. They are considered to represent a broader range of emotions, and the other one is the categorical model [7] that describes emotions in terms of a discrete set of basic emotions is still the most popular model for implementing FER because it can be directly related to intuitive definition of facial expressions. According to the feature representations, FER systems can be divided into two main categories: static image FER and dynamic sequence FER. In static-based methods [8],[9], the features information is encoded with only spatial information from the current single image, whereas dynamics-based methods [10] consider the temporal relation among contiguous frames in the input facial expression sequence. Based on these two vision-based methods, other modalities, such as audio and physiological channels, have also been used in multimodal systems [11] to assist the recognition of expression.

The majority of the traditional methods have used handcrafted features or shallow learning (e.g., texture/shape descriptors [12], local binary patterns (LBP)[13], and Gradient and Laplacian [14]) for FER. Recently, To achieve a good result for FER, a deep learning method based on CNN architecture has been widely used [15]. Since 2013, emotion recognition competitions such as FER2013 using Xceptionnet Architecture [16] and Emotion Recognition in the Wild (EmotiW) [17] have collected relatively sufficient training data from

challenging real-world scenarios, which implicitly promote the transition of FER from lab-controlled to in-the-wild settings. Meanwhile, due to the advancement in chip processing abilities (e.g., GPU units) and well-designed network architecture, studies in various fields transfer it to deep learning methods, which have achieved state-of-the-art recognition accuracy and exceeded previous results by a large margin[18]. Likewise, to effectively train data of facial expression, deep learning techniques have increasingly been implemented on different platforms like Tensorflow[19], Colab, and PyTorch.

In the past, for AU Detection on CK+ Dataset [20], few traditional BGCS [21] and HRBM [22] methods for detection of 13 Facial AUs were tried along with those deep learning methods such as DSCMR [22] and JPML [24] for 11 Facial AUs Except for AU17 and AU23. Recently Jing et al. [25] proposed a computational efficient end-to-end training deep neural network (CEDNN) model that uses spatial attention maps based on different images with different architecture such as res-L3M6, res-L18M1, and res-L18M1. Out of that, res-L3M6 is shown a good result on CK+ [20], whereas res-L18M1 and res-L18M1 shown on DISFA+ Dataset [37].

A few methods are listed in the literature for MMI Dataset AU detection. Models for dynamic classification provide a more standard way to encode the representation of facial expressions. With a few exceptions, most of the dynamic approaches to the classification of facial expressions are based on variants of Dynamic Bayesian Networks (DBN) (e.g., Hidden Markov Models (HMM) and Conditional Random Fields (CRF)). For example, the use of a generalization of the linear-chain CRF model, a Hidden Conditional Random Field (H-CRF), and other sequence-based methods (SVM-SB, H-CORF, and VLS-CRF)[29] models vastly outperform the existing frame-based methods (PFFL [26], LPQ-TOP [27] and FFD [28] on the AU detection task on MMI datasets.

For DISFA Dataset[30], The recent ARL [31] uses various structure and texture information in hierarchical region learning for Facial AUs detection using attention maps in different local regions. Similarly, SRERL [32] designed an important loss function named adaptive cross-entropy for imbalanced data training based on training samples proportion. Facial landmarks have also been used in different ways to generate attention maps. It includes an attention mechanism, and [33] contains correlations of action units. For example, in EAC-Net [34], a single attention map is created, which is fixed for each image by combining all regions associated with action units. Corneanu et al. [35] proposed a deep structural inference network (DSIN). It is based on CNN architecture to extract features of the entire image and patches separately. Then using interconnected recursive structure inference units, final integrated features are sent to a set of, and the information is transmitted iteratively to infer the structure between AUs.

Jacob et al.[36] also, follow an attention-based approach to estimate attention maps using landmarks instead of other methods. That helps in accurately focussing on the ROI relevant to each action unit. Also, in contrast to the fixed attention maps in the EAC-Net model, they predicted the attention maps during inference time.

All the above methods depend on the accuracy of landmark detection, and the definition of the AU center is very complicated. Many of these methods learn the features of each AU separately and only retain the relationship between AUs through joint feature learning. This method often loses the spatial relationship of AUs.

Also, To extract useful facial features for AU detection, researchers have tried many methods based on geometric features on the DISFA+ dataset [37]. Facial landmarks have strong robustness in describing geometric changes. Many traditional research methods are based on facial landmarks or texture features near facial landmarks. Zhao et al. [38] proposed a deep region multi-label learning (DRML) network trained end-to-end. The identification framework divides the response map evenly into 8×8 regions in the region layer and updates the weight in each patch independently. These attention maps are multiplied with CNN

features to help focus on the regions of interest. AU R-CNN [39] uses ResNet-101 as the backbone model for action unit detection. JAA-Net [40] jointly estimates the location of landmarks and the presence of action units. Landmarks are used to compute the attention map for each action unit separately.

The existing research is mainly focused on seven basic emotions (happy, sad, fear, disgust, anger, surprise, and neutral). However, humans express many kinds of emotions, including compound emotion which has not been explored much due to its complexity.

. Hence, there is a need for a FER system that detect particular basic and compound facial expression based on facial AUs, which is proposed.

The main motivation for building FER System as recognition of facial expressions plays a major role in many automated system applications like

(1) Medical Field [41] [42],
(2) E-Education [43] [44],
(3) video games [45]
(4) Human-robot interaction [46]
(5) Distance Counselling and psychotherapy [47] and
(6) Human-computer interaction [48].

Various machine vision and machine learning algorithms have been introduced to recognize facial expressions; still, recognize facial expressions accurately.

In Facial emotion, many facial Action Units (AU's) are common, leading to ambiguity. Most FER solutions are too general, and cross-database validation demonstrates their lack of robustness. Hence, there is a need for an efficient and robust real-life FER system that detect Facial Action Units (AUs) to detect particular basic and compound facial expression

To correctly recognize basic and compound human emotion from the Facial action unit on CK+, MMI, DISFA, and DISFA+ dataset, we proposed CNN based different architecture for Facial Expression Recognition (FER) system based on the required dataset.

3. Proposed Methodology for Action Unit Detection

The proposed methodology is based on a Convolution Neural network-based deep learning framework. In this, we are using standard CNN Architectures such as VGG, Resnet50, and Xceptionnet, which are well known for Image classification applications. These Standard Networks have starting layers that are used to learn Low-level features such as edges, blobs, and colors, and the last layers learn task-specific features, in our case, its Action Units (AU's).

We are modifying the standard network by replacing the last layers to extract features specific to the dataset. In the database, we have a variety of images, but to learn features faster, we are preprocessing Images so only the required part of the face is captured using the Haar Cascade algorithm to detect the frontal face.

The training datasets consist of different facial action units (AU's). However, the distribution of the training data is usually uneven. For example, in the CK+ dataset, the number of training images of a few AU's is less, while the other training pictures range from more. The imbalance of training data distribution will harm network performance. To eliminate this negative effect, we selected Action Units (AU's) whose sufficient training images are available only those AU's are considered. Also, we performed data augmentation (shown in Figure 1) and DataGeneration operation to balance the data distribution and generate more training data.

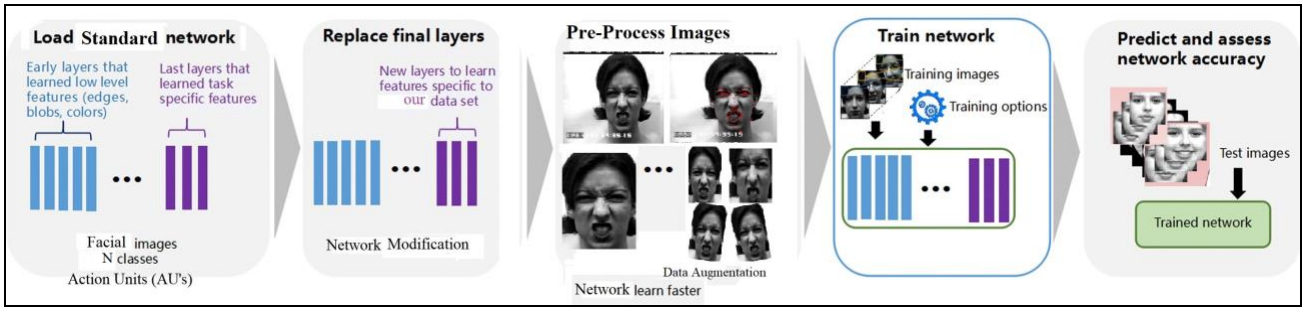

Figure 1: Proposed Methodology based on CNN based deep learning framework

To train, validate and test, network to improve performance, additional images are generated using Image DataGenerator, which Generates batches of tensor image data with real-time data augmentation. While training the network, training and validation Precision, Recall, and loss is calculated to get an idea of how well the network is trained. Finally, around 1000 test images are supplied for checking the model's performance, and various parameters such as precision, recall, accuracy, weighted f1-score, and hamming loss are calculated.

Multi-label classification

A multi-label classification problem involves mapping each sample in a dataset to a set of class labels. In this type of classification problem, the labels are not mutually exclusive. For example, when mentioning Action Units (AU's) for a given image, a given image might have multiple AUs

Because the labels are not mutually exclusive, the predictions and true labels are now vectors of label sets rather than vectors of labels. Multi-label metrics, therefore, extend the fundamental ideas of precision, recall, etc., to operations on sets. For example, a true positive for a given class now occurs when that class exists in the predicted set, in the true label set, for a specific data point.

Available metrics

Here we define set D of N images of a database as

$$D = \{d_0, d_1, ..., d_N\}$$

We define L_0, L_1,..., L_{N-1} to be a family of AU's label(L) sets and P_0, P_1,..., P_{N-1} to be a family of prediction (P) sets where L_i and P_i are the label set and prediction set, respectively, that correspond to Image d_i. The set of all unique labels is given by

$$L = \bigcup_{k=0}^{N-1} L_k$$

Similarly, Table 2 Shows a mathematical representation of Precision, Recall, Accuracy, and F1-score used in the calculation for every class.

Metric	Definition
Precision	$\frac{1}{N}\sum_{i=0}^{N-1}\frac{\lvert P_i \cap L_i\rvert}{\lvert P_i\rvert}$
Recall	$\frac{1}{N}\sum_{i=0}^{N-1}\frac{\lvert L_i \cap P_i\rvert}{\lvert L_i\rvert}$
Accuracy	$\frac{1}{N}\sum_{i=0}^{N-1}\frac{\lvert L_i \cap P_i\rvert}{\lvert L_i\rvert+\lvert P_i\rvert-\lvert L_i \cap P_i\rvert}$
F1 Measure	$\frac{1}{N}\sum_{i=0}^{N-1}2\frac{\lvert P_i \cap L_i\rvert}{\lvert P_i\rvert\cdot\lvert L_i\rvert}$

Table 2 Mathematical representation of Precision, Recall, Accuracy and F1 Measure

Loss Function

The use of cross-entropy for classification often gives different specific names based on the number of classes, mirroring the name of the classification task

Binary Cross-Entropy: Cross-entropy as a loss function for a binary classification task.

Categorical Cross-Entropy: Cross-entropy as a loss function for a multi-class classification task.

We are using categorical cross-entropy as a loss function.

We used a confusion matrix to evaluate the accuracy between actual and predicted labels.

Cohn-Kanade (CK+) database [20]

This dataset was introduced by Lucey et al. [20]. 210 persons, aged 18 to 50, have been recorded depicting emotions. Both female and male persons are present and from different backgrounds. 81% are Euro-Americans and 13% Afro-Americans, and 6% other groups. An experimenter instructed participants to perform a series of 23 facial displays; these included single action units and combinations of action units. Each video began and ended in a neutral face, with any exceptions noted. Image sequences for frontal pose and 30-degree views were digitized into either 640x490 or 640x480 pixel arrays with 8- bit grayscale or 24-bit color values. Figure 2 shows a few sample images of the CK+ Database.

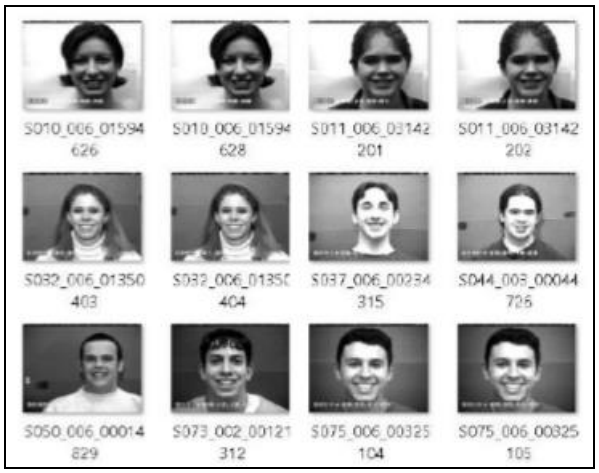

Figure 2: Sample Images of CK+ Database

VGG16 Modified Architecture for CK+ Database

VGG16 is a convolution neural net (CNN) architecture used to win the ILSVR (Imagenet) competition in 2014. It is considered one of the excellent vision model architecture to date. The unique thing about VGG16 is that instead of having a large number of hyper-parameter, they focused on having convolution layers of a 3x3 filter with a stride 1 and always used the same padding and max pool layer of 2x2 filter of stride 2. It consistently follows this arrangement of convolution and max pool layers throughout the architecture. In the end, it has 2 FC(fully connected layers) followed by a softmax for output. The 16 in VGG16 refers to it having 16 layers that have weights. This network is pretty extensive and has about 138 million (approx) parameters.

For our application of Facial Action unit detection, we modified its architecture slightly by changing a few final fully connected layers to get good results of detecting 17 Facial Action Units {1, 2, 4, 5, 6, 7, 9, 12, 14, 15, 17, 20, 23, 24, 25, 26, 27}. Figure 3a & 3b shows the detailed architecture and Proposed Methodology based on CNN-based deep learning for CK+ Database, which shows after-action unit detection is mapped based on [5] to find Basic and Compound emotion.

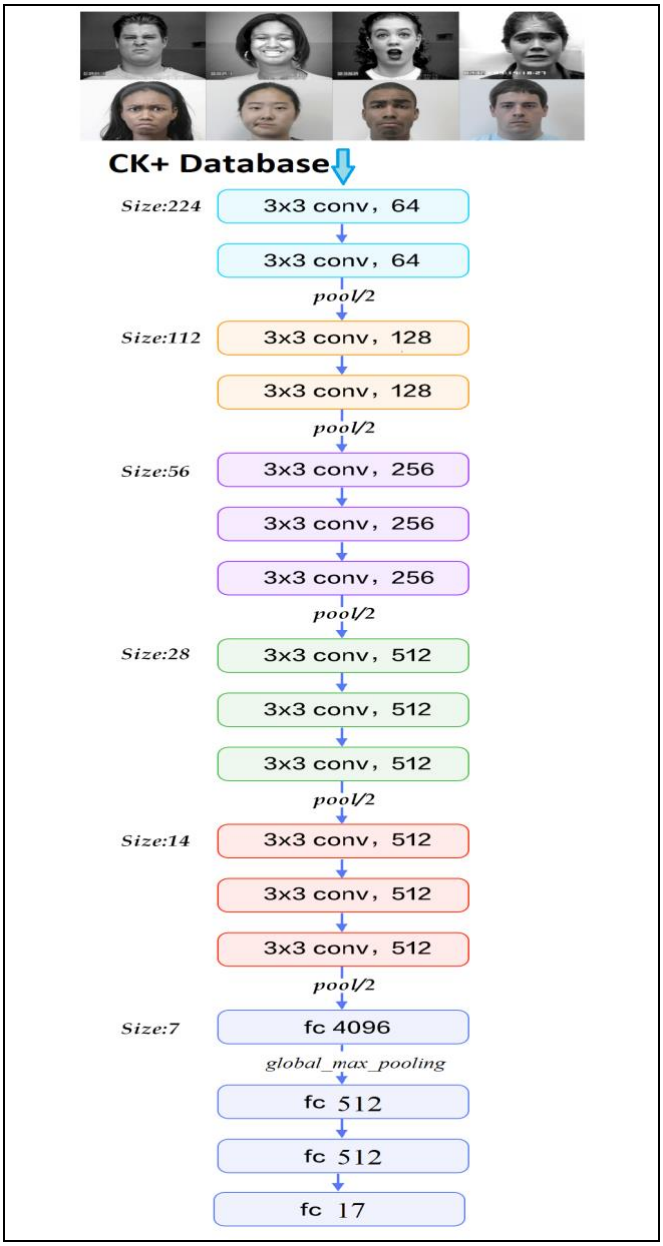

Figure 3a: Modified VGG16 architecture for CK+ Database

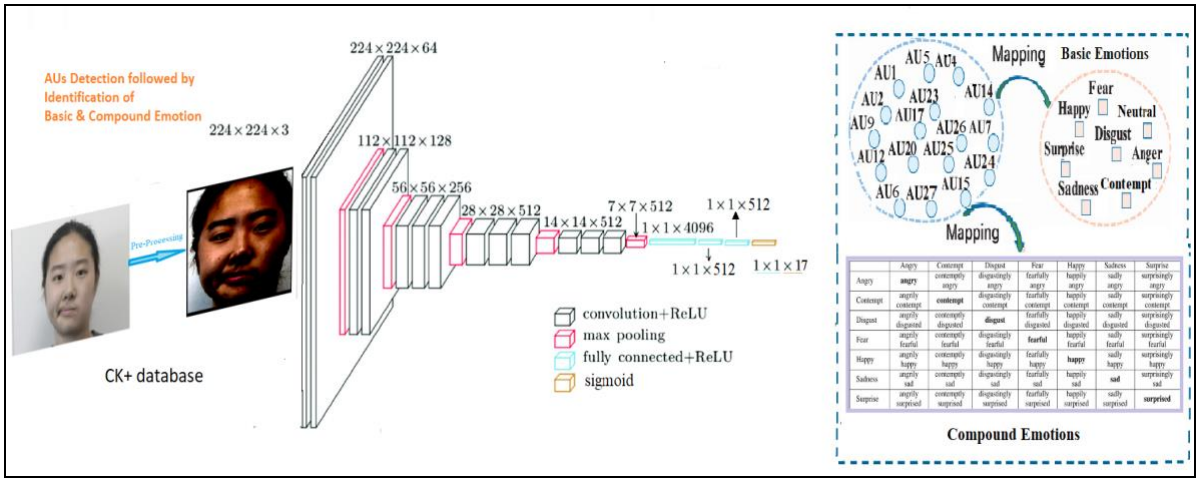

Figure 3b: Proposed Methodology based on CNN based deep learning for CK+ Database

MMI database [27]

The MMI dataset has been introduced by Pantic et al. [27], which consists of over 2900 videos and high-resolution still images of 75 subjects. It is fully annotated for AUs' presence in videos (event coding) and partially coded on frame-level, indicating whether an AU is in either the neutral, onset, apex or offset phase for each frame. Figure 4 shows a few sample images of the MMI Database.

Figure 4: Sample Images of MMI Database

Xception Modified Architecture for MMI Database

Xception [15] is a deep convolutional neural network architecture that involves Depthwise Separable Convolutions. Google researchers presented an interpretation of Inception modules in CNNs as an intermediate step between regular convolution and the depthwise separable convolution operation (a depthwise convolution followed by a pointwise convolution). This observation leads them to propose a novel deep convolutional neural network architecture inspired by Inception, where Inception modules have been replaced with depthwise separable convolutions.

The data first goes through the entry flow, then through the middle flow, which is repeated eight times, and finally through the exit flow. Note that all Convolution and Separable Convolution layers are followed by batch normalization.

For our application of Facial Action unit detection we modified its architecture slightly by changing few final fully connected layers to get good results of detecting 16 Facial Action Units {1,2,4,6,7, 9,10,11,12,14,15,17,20,24,25,26}. Figure 5 shows the detailed architecture and Proposed Methodology based on CNN-based deep learning for MMI Database, after-action unit detection using the multi-label classification method. After that, it's mapped based on [5] to respective Basic and Compound emotions.

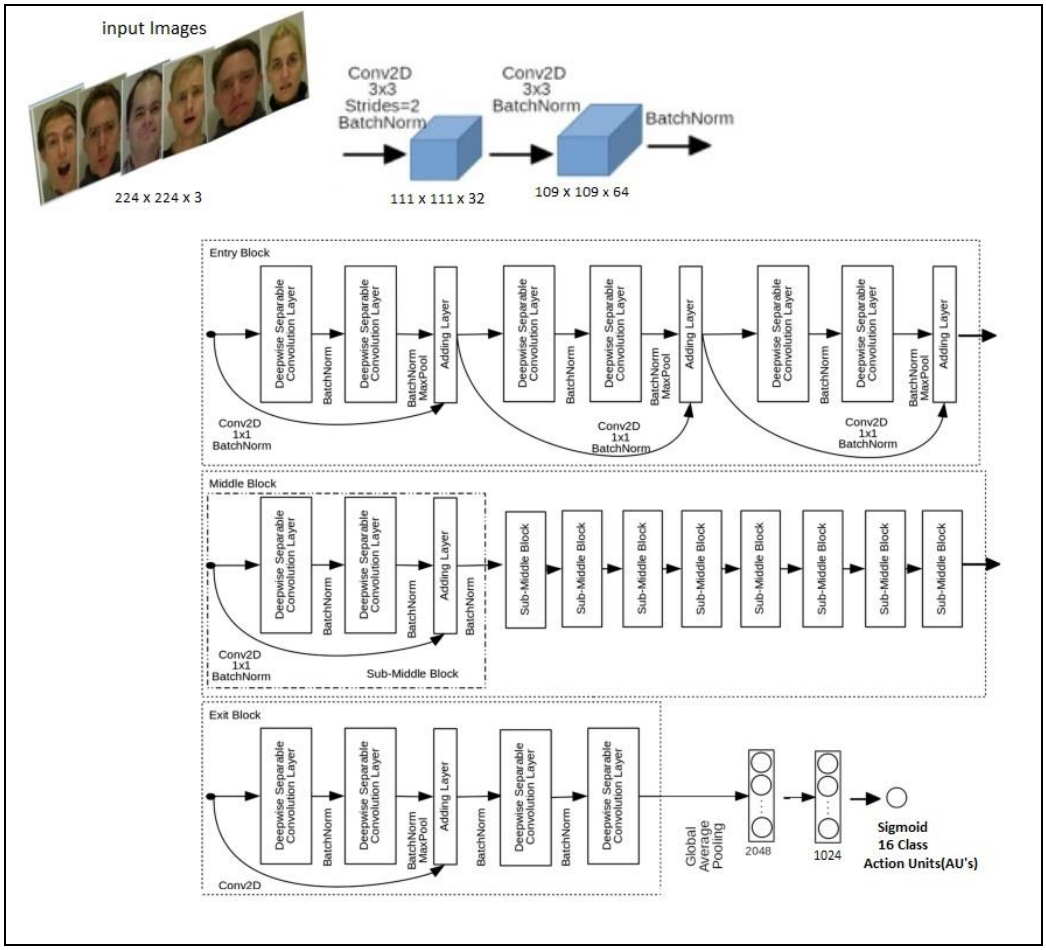

Figure 5: A proposed framework based on CNN based deep learning for MMI Database

DISFA Database [30]

The Denver Intensity of Spontaneous Facial Action Database is a non-posed facial expression database for automatic action unit detection and its intensities described by FACS. This database contains stereo videos of 27 adult subjects (12 females and 15 males) with different ethnicities. The images were acquired using PtGrey stereo imaging system at high resolution (1024×768). Two human FACS experts manually scored the intensity of AU's (0-5 scale) for all video frames. The database also includes 66 facial landmark points of each image. Figure 6 shows a few sample images of the DISFA Database.

Figure 6: Sample Images of DISFA Database

Xception [15] Modified Architecture for DISFA Database

For the DISFA dataset, the same modified Xception architecture is used for our Facial Action unit detection application. We modified its architecture slightly by changing a few

final fully connected layers to get good results of detecting 12 Facial Action Units {1,2,4,5,6,9,12,15,17,20,25,26}. Figure 7 shows the detailed architecture and Proposed Methodology based on CNN-based deep learning for DISFA Database, which shows after-action unit detection is mapped based on [5] to find Basic and Compound emotion.

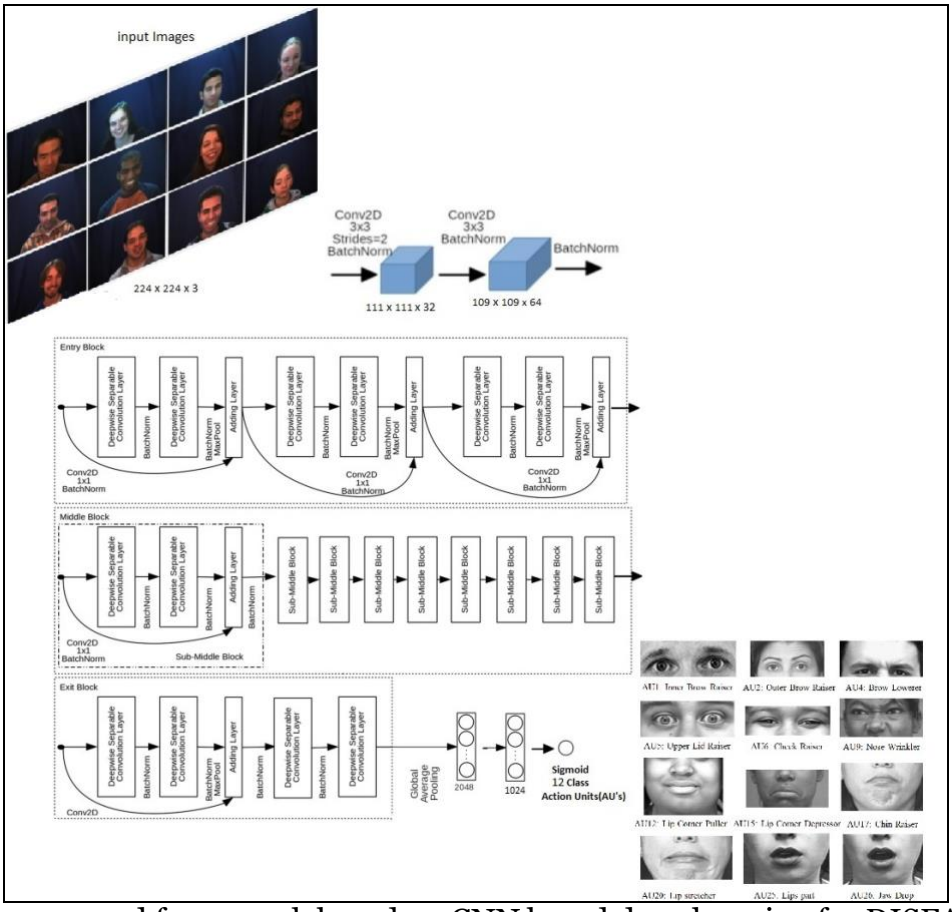

Figure 7: A proposed framework based on CNN based deep learning for DISFA Database

Extended Denver Intensity of Spontaneous Facial Action Database (DISFA+)[37]

DISFA+ The Extended Denver Intensity of Spontaneous Facial Action Database extends DISFA. DISFA+ has the following features:

a. it contains a large set of posed and non-posed facial expressions data for the same group of individuals,

b. it provides the manually labeled frame-based annotations of the 5-level intensity of twelve FACS facial actions,

c. it provides metadata (i.e., facial landmark points in addition to the self-report of each individual regarding every posed facial expression)

Figure 8 shows a few sample images of the DISFA+ Database.

Figure 8: Sample Images of DISFA+ Database

The framework of the proposed method for DISFA+ Database

ResNet50 [15] is a variant of the ResNet model, which has 48 Convolution layers and 1 MaxPool, and 1 Average Pool layer. It has 3.8 x 10^9 Floating points operations. It is a widely used ResNet model. In 2012, at the LSVRC2012 classification contest, AlexNet won the first prize. After that, ResNet was the most exciting thing about computer vision and the deep learning world.

Because of the framework that ResNets presented, it was made possible to train ultra-deep neural networks, and by that mean, a network can contain hundreds or thousands of layers and still achieve outstanding performance.

For our application of Facial Action unit detection, we modified its architecture slightly by changing a few final fully connected layers to get good results of detecting 12 Facial Action Units {1,2,4,5,6,9,12,15,17,20,25,26}. Figure 9 shows the detailed architecture and Proposed Methodology based on CNN-based deep learning for DISFA+ Database, which shows after-action unit detection is mapped based on [5] to find Basic and Compound emotion.

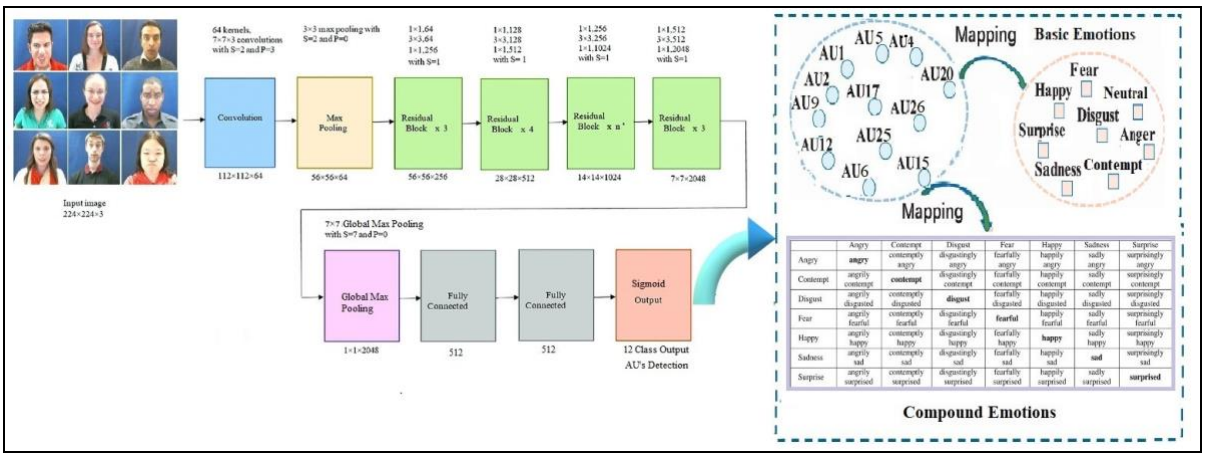

Figure 9: A proposed framework based on CNN based deep learning for DISFA+ Database

4. Results

Multi-label Confusion Matrix

The multi-label confusion matrix function computes class-wise or sample-wise multi-label confusion matrix to evaluate the accuracy of classification. Multi-label confusion matrix also treats multi-class data as if it were multi-label, as this is a transformation commonly applied to evaluate multi-class problems with binary classification metrics (such as precision, recall, etc.).

We use a confusion matrix to evaluate the accuracy between actual and predicted labels.

Confusion Matrix of CK+ Database

CK+ Database has 17 Facial Action Units {1, 2, 4, 5, 6, 7, 9, 12, 14, 15, 17, 20, 23, 24, 25, 26, 27} and also we added 0 as if model fails to detect particular Action Units then it is considered as 0. Confusion Matrix shows that most classes are classified with good accuracy except AU 14, 20, 23 & 24, which are related to cheek & lips movement. Figure 10 shows Confusion Matrix for CK+ Database.

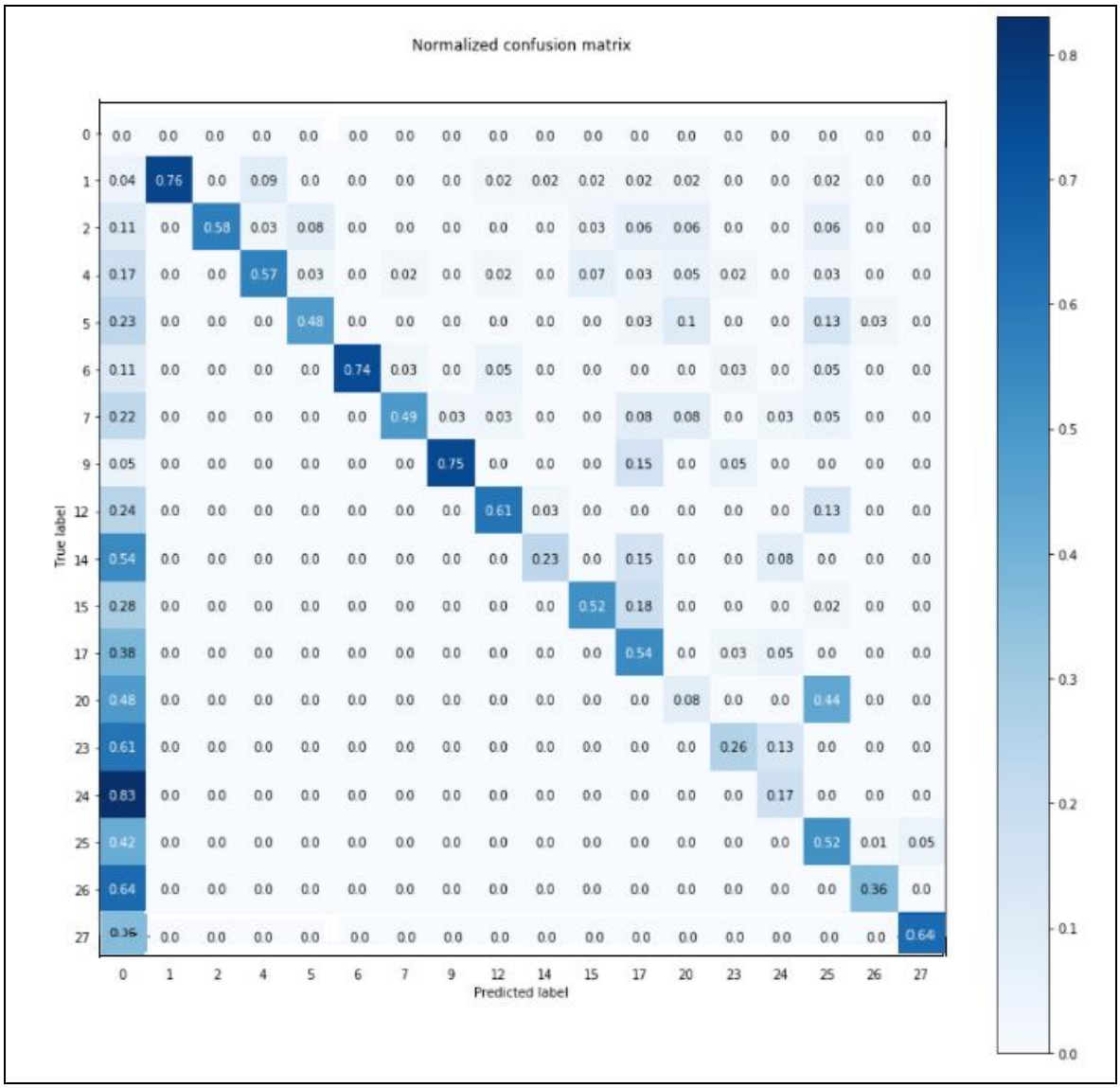

Figure 10: Confusion Matrix for CK+ Database

Confusion Matrix of MMI Database

MMI Database has 16 Facial Action Units {1,2,4,6,7, 9,10,11,12,14,15,17,20,24,25,26} and also we added 0 as if model fails to detect particular Action Units then it is considered as 0. Confusion Matrix shows most of the classes are classified which good accuracy except AU 17 & 24 which are related to chin & lips movement. Figure 11 shows Confusion Matrix for MMI Database.

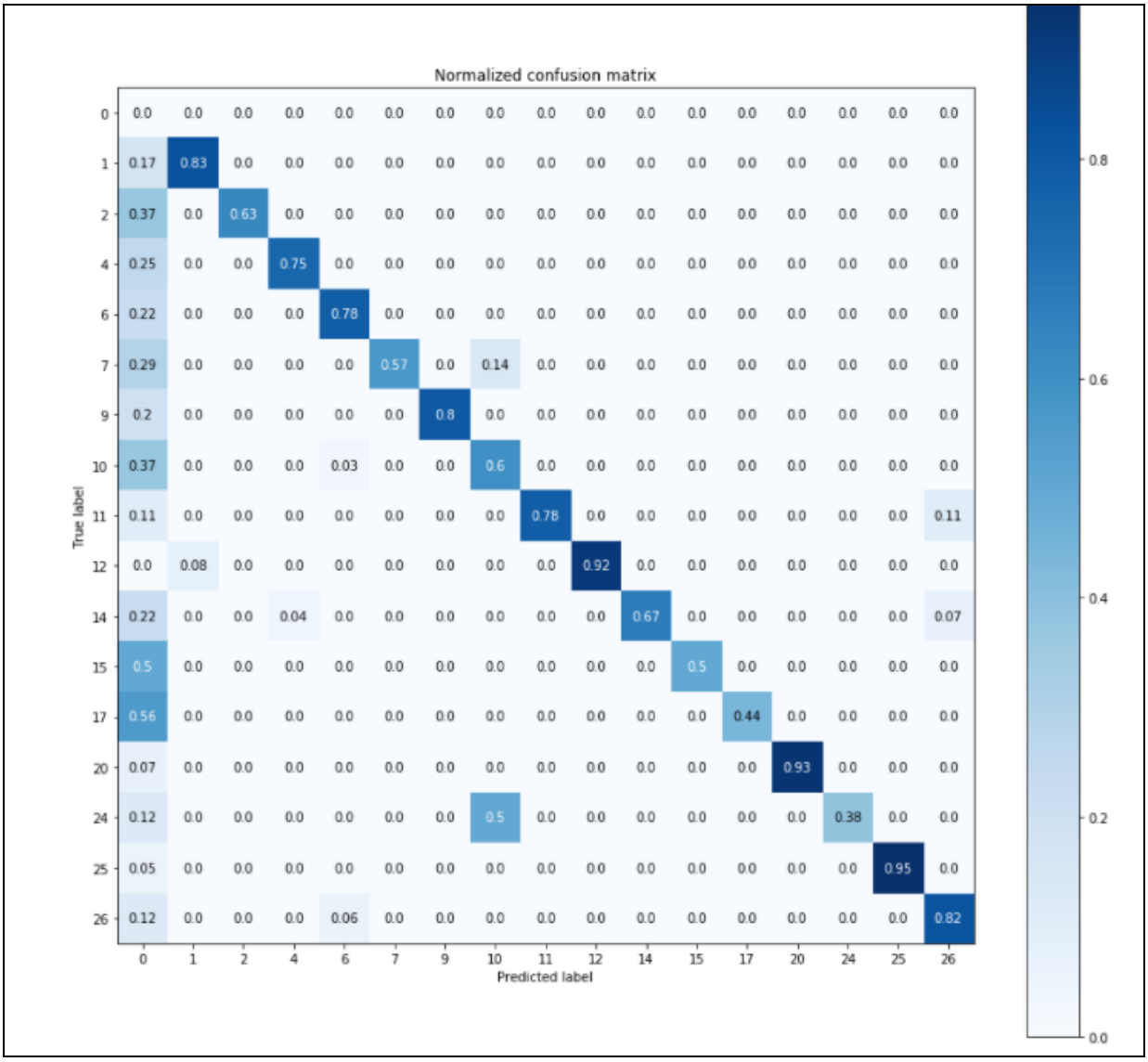

Figure 11: Confusion Matrix for MMI Database

Confusion Matrix of DISFA Database

DISFA Database has 12 Facial Action Units {1,2,4,5,6,9,12,15,17,20,25,26} and also we added 0 as if model fails to detect particular Action Units then it is considered as 0. Confusion Matrix shows that most classes are classified with good accuracy except AU 9, related to Nose wrinkle. Figure 12 shows the Confusion Matrix for DISFA Database.

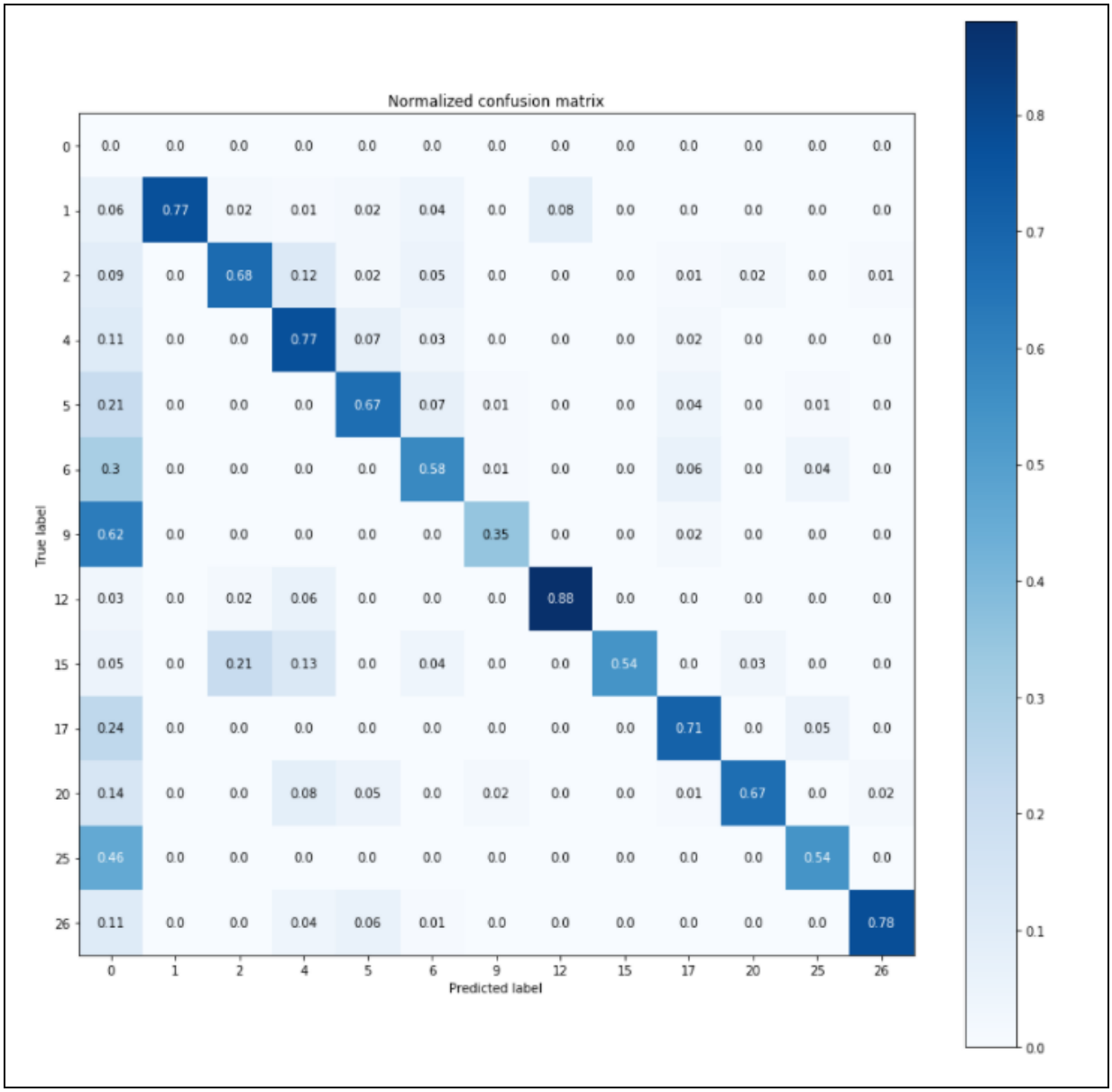

Figure 12: Confusion Matrix for DISFA Database

Confusion Matrix of DISFA+ Database

DISFA+ Database has 12 Facial Action Units {1,2,4,5,6,9,12,15,17,20,25,26} and also we added 0 as if model fails to detect particular Action Units then it is considered as 0. Confusion Matrix shows that most classes are classified with good accuracy except AU 9, related to Nose wrinkle. Figure 13 shows the Confusion Matrix for DISFA+ Database.

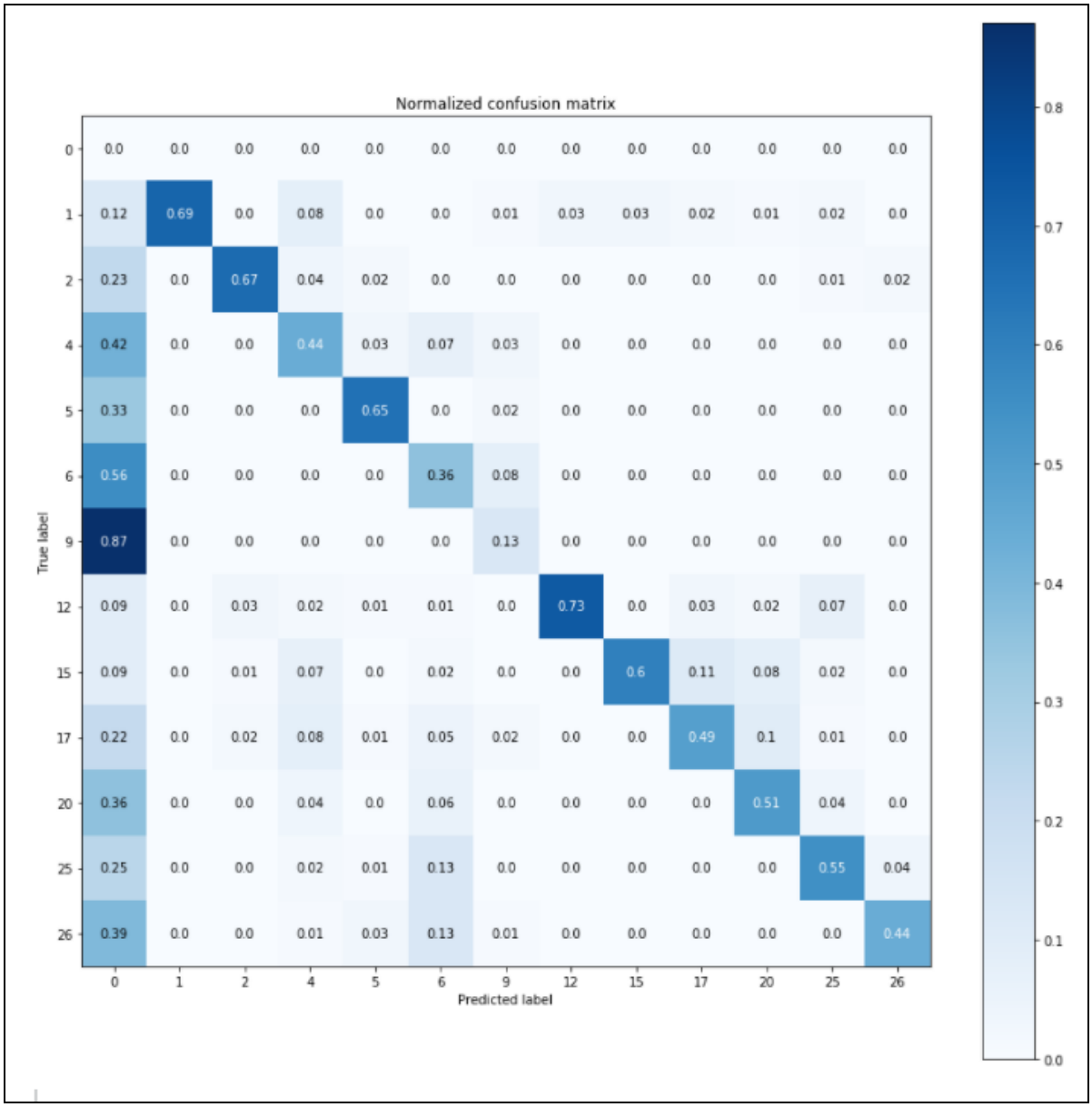

Figure 13: Confusion Matrix for DISFA+ Database

Summary result of precision, recall, and f1-score

To summarise, the precision, recall, and f1-score are plotted for all available Facial AUs in the same figure to give a good understanding of the performance of the proposed model.

Figure 14-17 shows the Summary result of precision, recall, and f1-score concerning Facial AU's for CK+, MMI, DISFA, and DISFA+ Dataset.

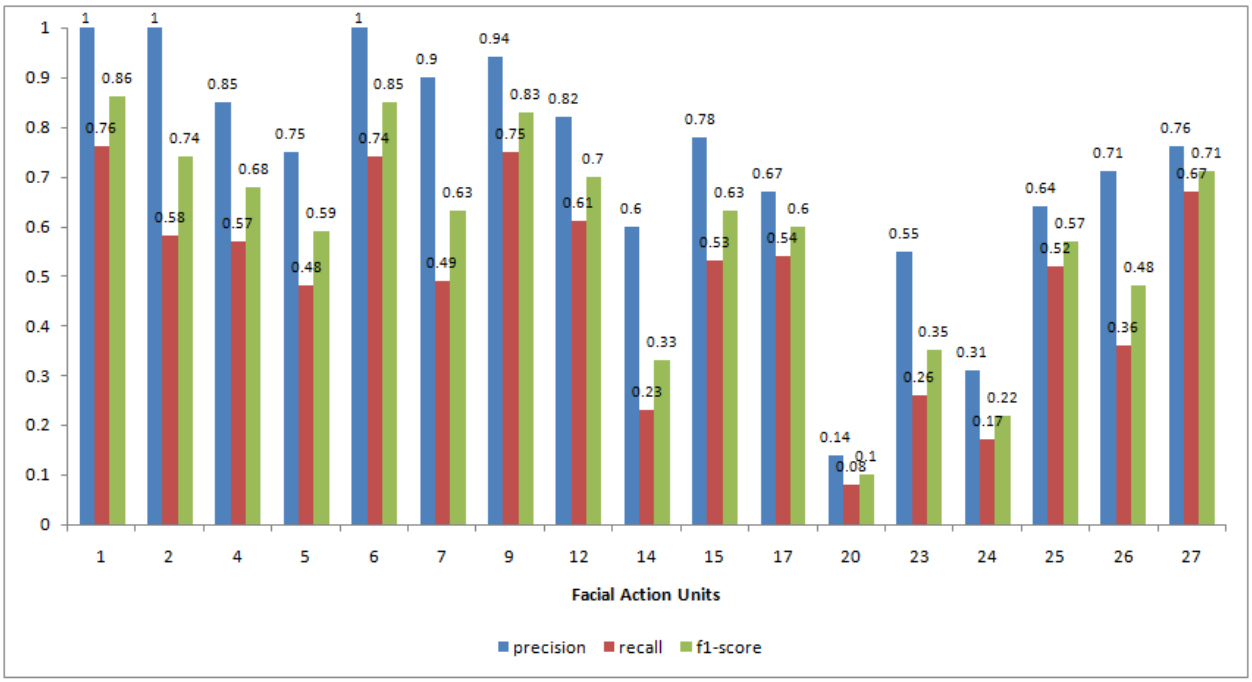

Figure 14: The summary result of precision, recall, and f1-score for CK+ Database

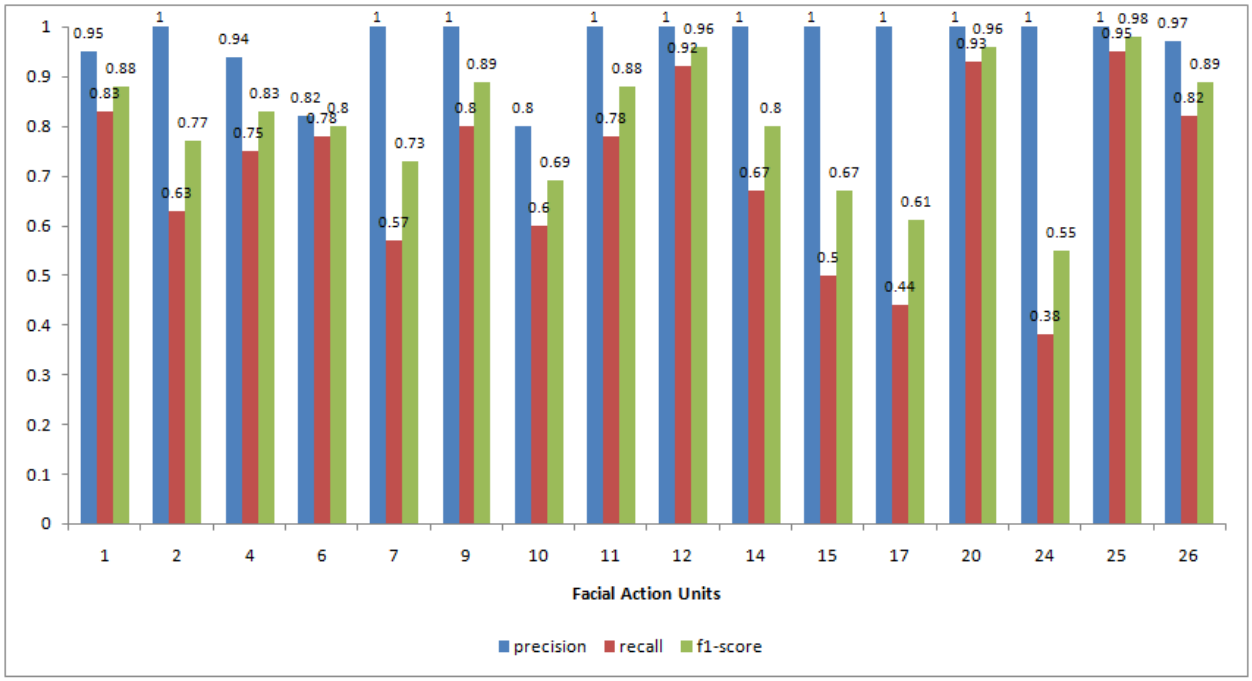

Figure 15: The summary result of precision, recall, and f1-score for MMI Database

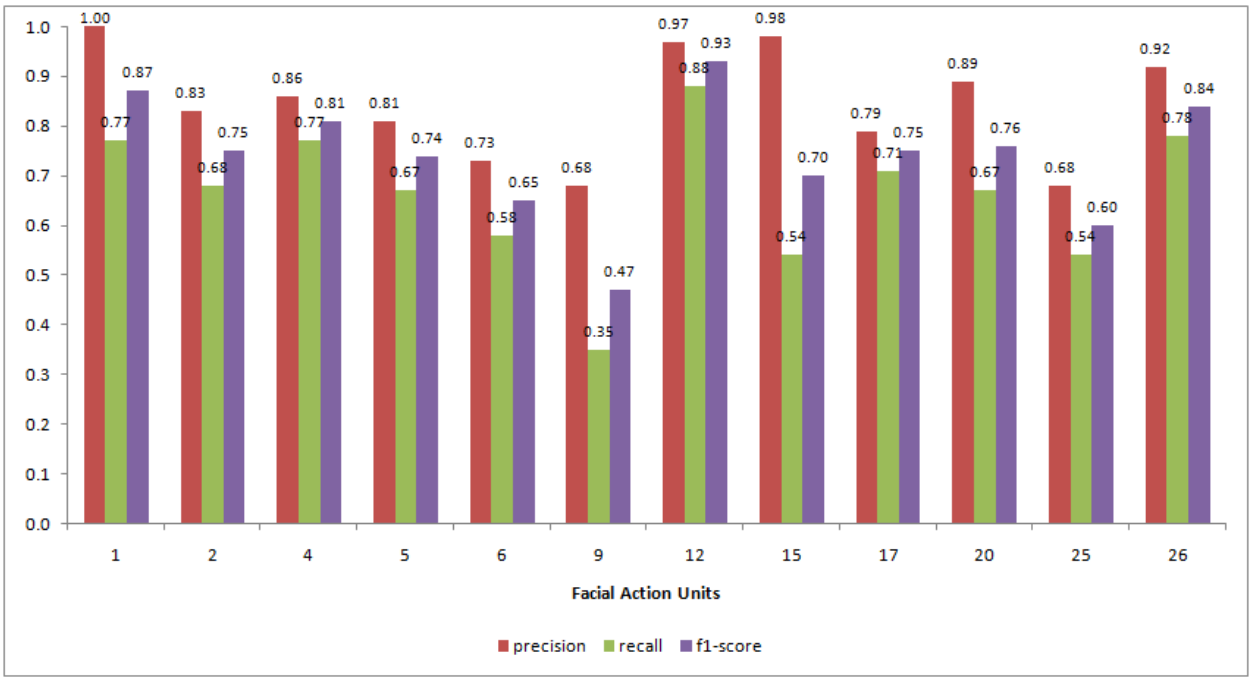

Figure 16: The summary result of precision, recall, and f1-score for DISFA Database

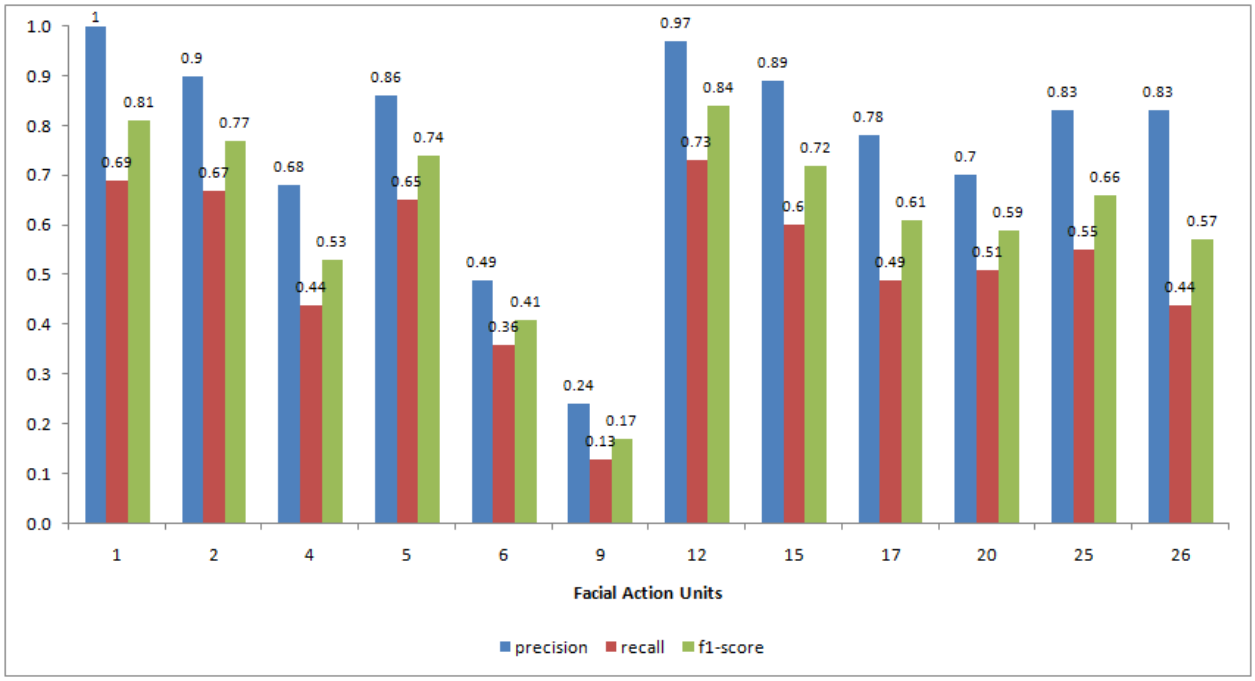

Figure 17: The summary result of precision, recall, and f1-score for DISFA+ Database

Action Unit Detection and Emotion Analysis on Sample Test Image

Action Unit Detection and Emotion Analysis of Sample test Images is done at the end to check the model's working from CK+, MMI, DISFA, and DISFA+ Dataset. It shows True Action Units (AU's) given in the database Predicted AU's by the proposed model and finally Basic and Compound Emotion Analysis by mapping [5].

Figure 18-22 shows Action Unit Detection and Emotion Analysis of Sample test Images to show a variety of combinations of AU's and Emotions from CK+, MMI, DISFA, and DISFA+ Dataset, respectively. And it gives a good understanding of Model accuracy, indicating actual

and predicted AU's for correct emotion, showing the accurate classification of Basic and Compound emotion through mapping.

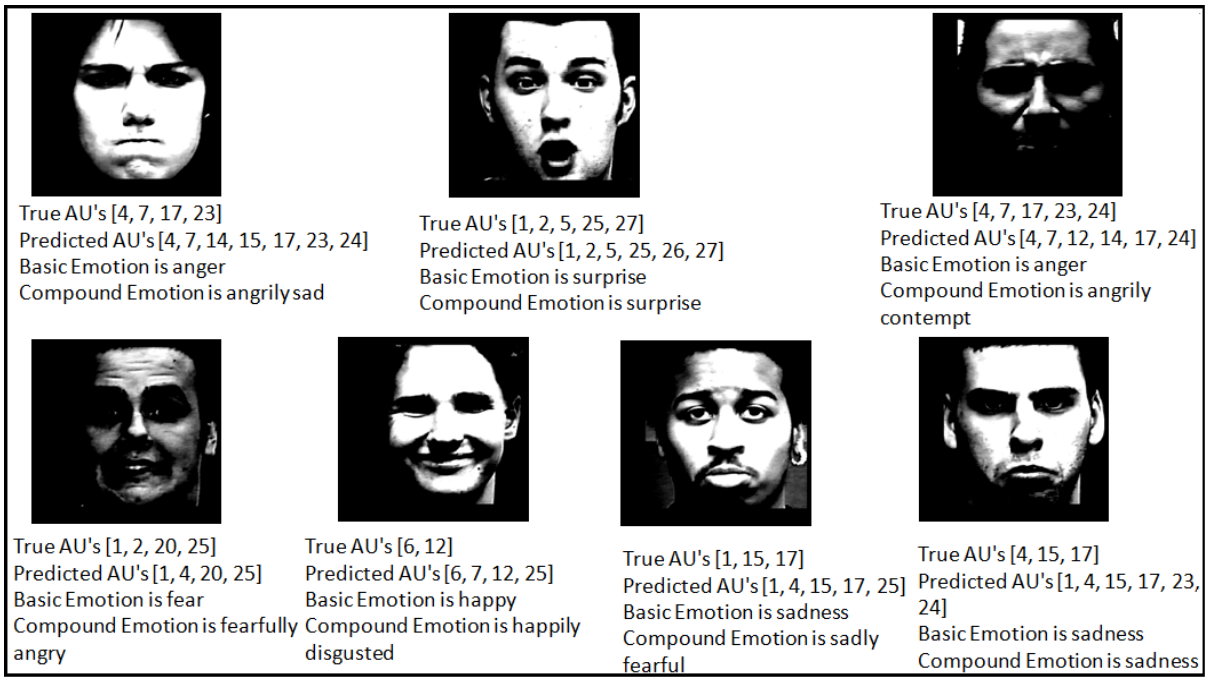

Figure 18: Results of CK+ Database

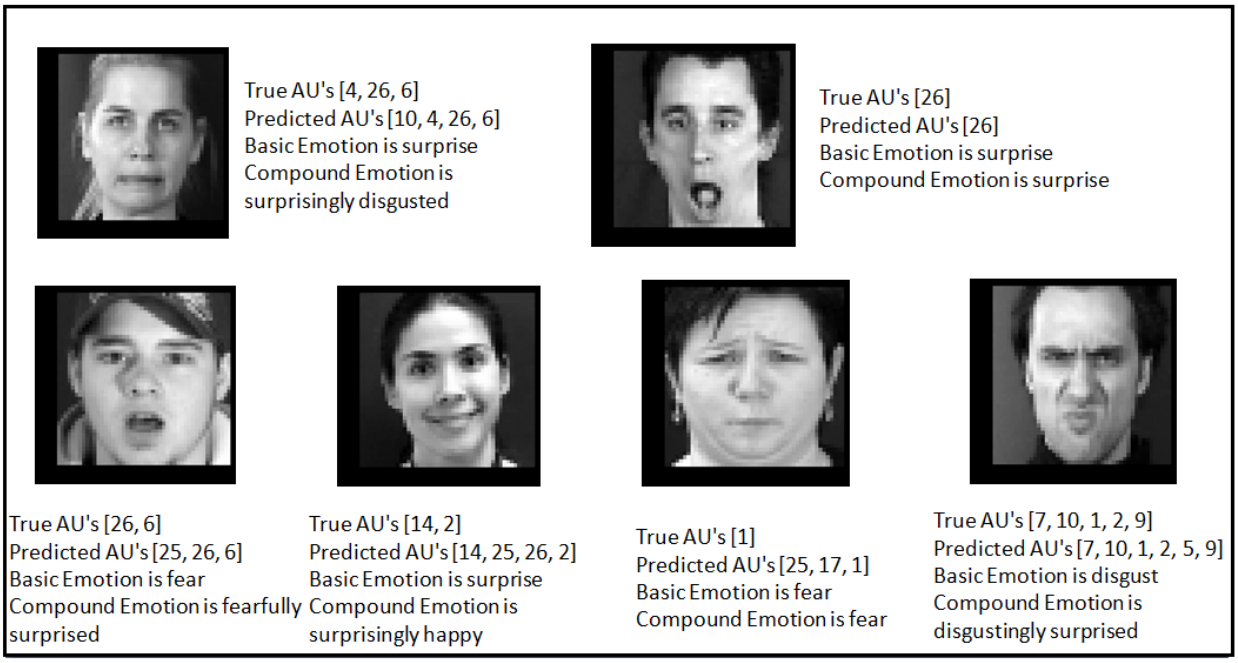

Figure 19: Results of MMI Database

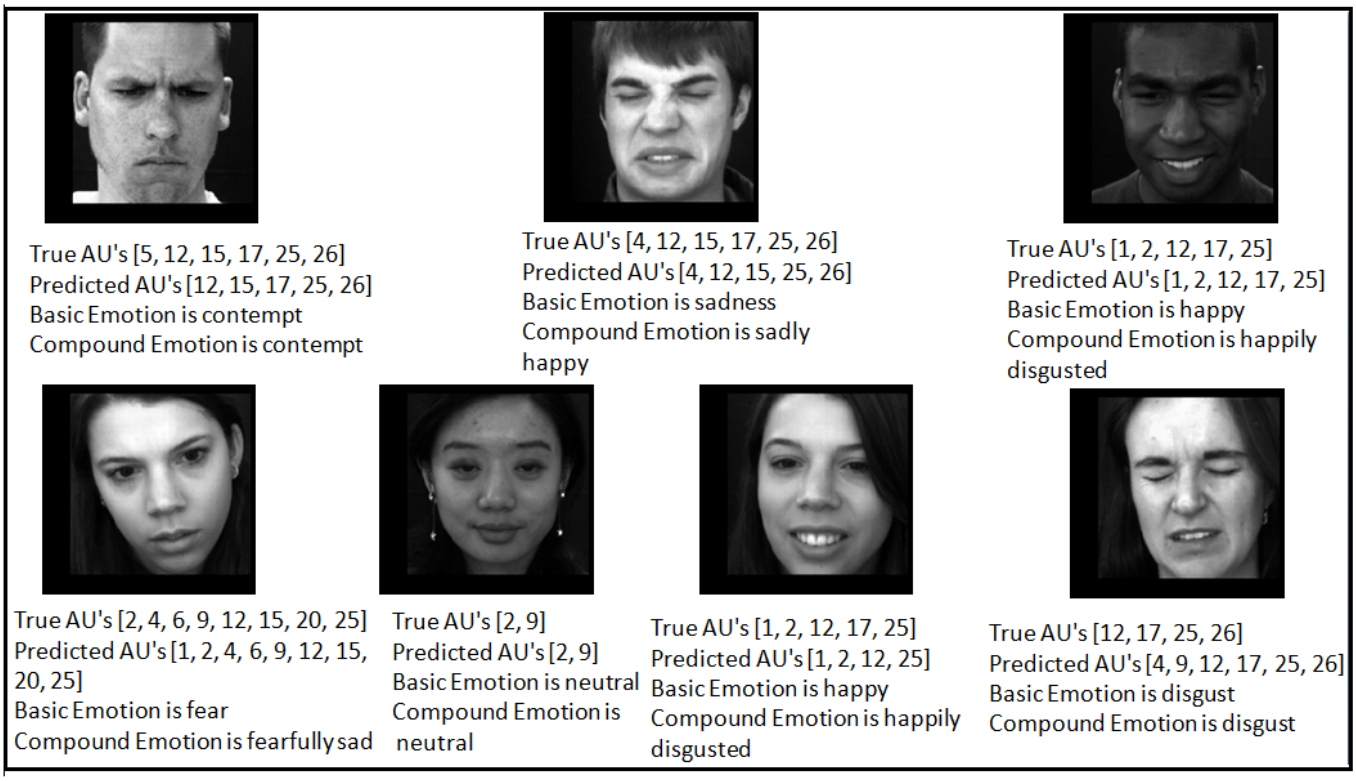

Figure 20: Results of DISFA Database

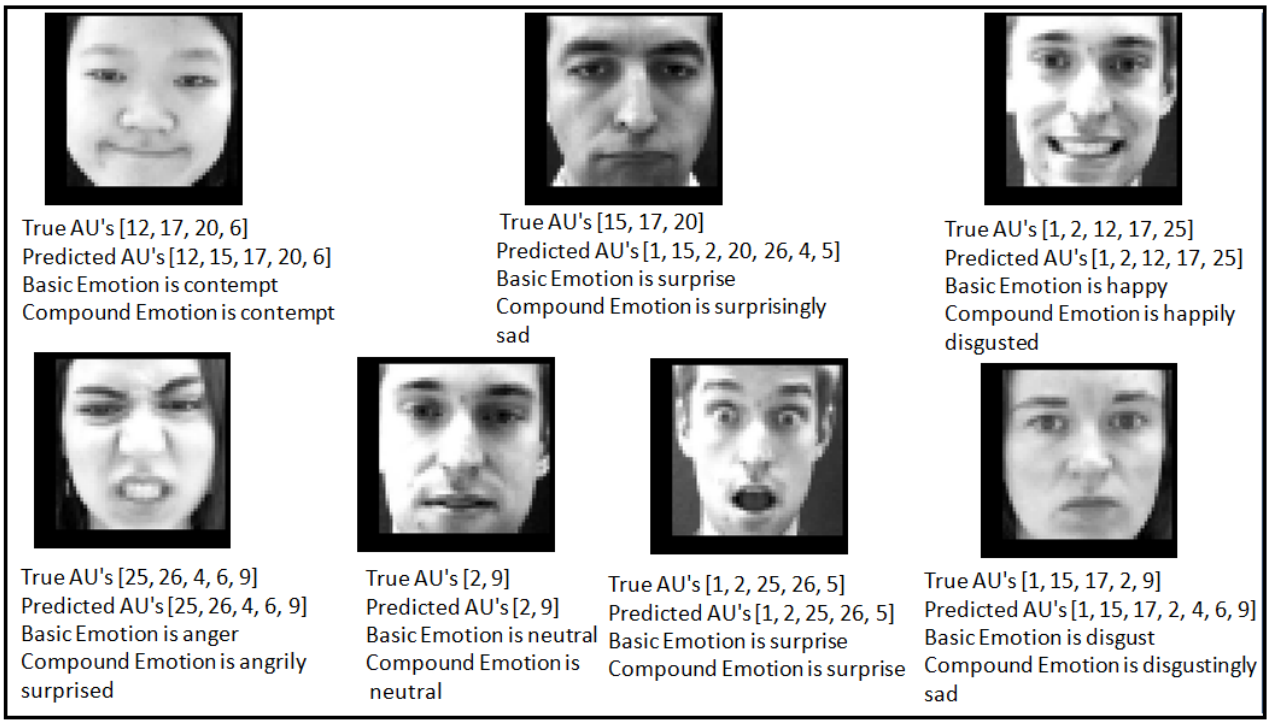

Figure 21: Results of DISFA+ Database

Comparison with State-of-the-Art Methods

We compare our method against state-of-the-art methods. For AU detection On the CK+ dataset, we compare our model with the recent related work taken from BGCS[16], HRBM[18], JPML[15], DSCMR[22], and res-L3M6[25]. Table 2 shows a comparison with

state-of-the-art results In terms of F1-score, and it shows the proposed method performs best for AUs 6 & 7.

Table 2: state of the art results comparison for AU Detection on CK+ Database using the F1-score metric (higher is better).

AU	BGCS	HRBM	res-L3M6	JPML	DSCMR	L-DSCMR	res-L3M6 (12 classes)[25]	Our Model
1	71	86.9	85.9	50	54.2	64.3	**87.5**	86
2	66.9	85.5	**88.4**	40	64.7	58.6	86.1	74
4	63.6	72.6	80.7	72.3	61.1	69.1	**82.1**	68
5	59.6	72	74.2	52.6	42.2	66.7	**74.9**	59
6	57.8	61.7	70	57.8	68.2	62.2	68.1	**85**
7	46.6	54.5	61.6	24.4	35.9	47.1	55.8	**63**
9	51.9	85.9	89.2	54.5	53.8	51.5	**89.9**	83
12	73.4	72.7	87.2	75.6	80	75.6	[87.3]	70
14	-	-	-	-	-	-	-	**33**
15	-	-	-	-	-	-	-	**63**
17	76.5	81.7	86.4	81.9	**89.7**	82.4	84.6	60
20	-	-	-	-	-	-	-	**10**
23	22.2	56.6	45.8	41.7	**75.0**	73.3	32.6	35
24	23.9	35.3	**46.2**	30.8	36.4	26.7	43.2	22
25	78.5	92.6	**93.6**	76.4	85.5	89.1	91.4	57
26	-	-	-	-	-	-	-	**48**
27	57.9	87.7	**89.6**	-	-	-	-	71
Avg	57.7	72.7	**76.8**	54.8	62.2	63.9	73.6	62

We compared the result with the existing reported frame-based methods (PFFL [26], LPQ-TOP [27], FFD [28]) and the sequence-based methods (SVM-SB, H-CRF, H-CORF, and VLS-CRF)[29] on the AU detection task. Yet, the proposed method outperforms these models on the MMI dataset. Table 3 shows a comparison with state-of-the-art results In terms of F1-score (Accuracy), and it shows the proposed method performs best for AU 4.

Table 3 State of art results comparison for AU Detection on MMI The numbers shown represent the F-1 (Accuracy) scores in % for each method.

AU	PFFL [26]	FFD [28]	LPQ-TOP[27]	SVM-SB [29]	H-CRF [29]	H-COR [29]	F VSL-CRF [29]	Ours Model
1	70.0 (—)	72.7 (—)	85.6 (—)	81.8 (85.9)	87.4 (87.5)	87.5 (87.5)	91.7 (91.2)	88 (95)
2	62.3 (—)	72.7 (—)	79.4 (—)	79.6 (79.8)	87.4 (87.5)	78.5 (78.4)	87.5 (87.8)	77 (100)
4	64.0 (—)	67.3 (—)	81.2 (—)	58.8 (61.6)	57.0 (57.1)	63.8 (64.4)	69.5 (70.7)	83 (94)
6	63.4 (—)	73.7 (—)	87.2 (—)	63.9 (68.2)	67.0 (67.3)	63.3 (63.5)	78.7 (78.8)	80 (82)
7	39.2 (—)	36.4 (—)	80.9 (—)	77.9 (84.0)	66.1 (66.6)	86.7 (86.6)	87.7 (87.6)	73 (100)
9	NA	NA	NA	NA	NA	NA	NA	89 (100)
10	NA	NA	NA	NA	NA	NA	NA	69 (80)
11	NA	NA	NA	NA	NA	NA	NA	88 (100)
12	NA	NA	NA	NA	NA	NA	NA	96 (100)
14	NA	NA	NA	NA	NA	NA	NA	80 (100)
15	NA	NA	NA	NA	NA	NA	NA	67 (100)
17	NA	NA	NA	NA	NA	NA	NA	61 (100)

20	NA	NA	NA	NA	NA	NA	NA	96 (100)
24	NA	NA	NA	NA	NA	NA	NA	55 (100)
25	NA	NA	NA	NA	NA	NA	NA	98 (100)
26	NA	NA	NA	NA	NA	NA	NA	89 (97)
AVG	59.8 (—)	64.6 (—)	82.6 (—)	72.4 (75.9)	73.0 (73.2)	76.0 (76.1)	83.0 (83.2)	72 (94)

For AU detection On the DISFA dataset[30], The performance is evaluated for 12 action units, we compare our model with the recent related work taken from the papers ARL[31], SRERL[32], LP [33], JPML [24], EAC [34], DSIN [35] and Jacob [36]. Table 4 shows a comparison with state-of-the-art results In terms of F1-score, and it shows the proposed method performs best on all of the Aus except AU 9 & 15.

Table 4: state of the art results comparison for AU Detection on the DISFA Database using the F1-score metric (higher is better).

AU	**DSIN [20]**	**LP[9]**	**SRERL[3]**	**EAC[17]**	**JAA[18]**	**ARL[19]**	**Jocob [21]**	**Ours Model**
1	42.4	29.9	45.7	41.5	43.7	43.9	46.1	**87**
2	39.0	24.7	47.8	26.4	46.2	42.1	48.6	**75**
4	68.4	72.7	59.6	66.4	56.0	63.6	72.8	**81**
5	NA	NA	NA	NA	NA	NA	NA	**74**
6	28.6	46.8	47.1	50.7	41.4	41.8	56.7	**65**
9	46.8	49.6	45.6	**80.5**	44.7	40.0	50.0	47
12	70.8	72.9	73.5	89.3	69.6	76.2	72.1	**93**
15	NA	NA	NA	NA	NA	NA	NA	**70**
17	NA	NA	NA	NA	NA	NA	NA	**75**
20	NA	NA	NA	NA	NA	NA	NA	**76**
25	90.4	93.8	84.3	88.9	88.3	**95.2**	90.8	60
26	42.2	65.0	43.6	15.6	58.4	66.8	55.4	**81**

To validate the effectiveness of our method, we compare our method with the related methods such as DRML[38], AU R-CNN[39], JAA-Net[40], res-L18M1, and res-L18M1 (bt24)[25] bold numbers indicate the best score on DISFA+ ˆ dataset[37]

The experimental results on the DISFA+ dataset are shown in Table 5. Our model is superior for AU 15 & 20 compared with the state-of-the-art works. There is severe data imbalance in DISFA+, which results in the performance of different AUs oscillating seriously.

Table 5: state of the art results comparison for AU Detection on DISFA+ Database using the F1-score metric (higher is better).

AU	**DRML[38]**	**AU R-CNN[39]**	**JAA-Net [40]**	**res-L18M1 [25]**	**res-L18M1(bt24)[25]**	**Ours Model**
1	27.3	47.8	**83.9**	83.2	82.3	81.0
2	22.2	42.7	80.5	80.1	**80.7**	77.0
4	51	55.7	79.3	78.4	**79.7**	53.0
5	36.4	47.8	**78.4**	74.3	76.5	74.0
6	56.2	42.7	78.2	**82.3**	80.6	41.0
9	32	24.3	67.6	**74.7**	69.5	17.0
12	38.5	47.4	**84.6**	83.8	83.6	84.0
15	22.7	23.5	55.4	55.9	55.1	**72.0**
17	27.1	6.4	60.4	**65**	64.8	61.0
20	16.3	28.7	48.5	49.7	51.8	**59.0**

25	56.5	53.1	85.1	88.2	**89.9**	66.0
26	42.3	38.6	69	**76.6**	74.1	57.0
Avg	35.7	38.2	72.6	**74.4**	74	61.8

Finally, we can summarize the result of the CK+, MMI, DISFA, and DISFA+ Dataset with state-of-the-art results compared. It shows that model based on Xception architecture is giving more promising results than VGG and ResNet-based architecture.

5. Conclusion

This research aimed to present different CNN-based architecture with modified standard architecture such as VGG and Xceptionnet and to detect facial action units on CK+, MMI, DISFA, and DISFA+ Dataset more precisely and interns used to map with basic and compound human emotion. The proposed method gives superior accuracy for detecting most Action units on all datasets compared with state-of-the-art results. For Action Unit detection overall accuracy of the proposed Xceptionnet network for MMI & DISFA are giving promising results average F1 score is 72% and 74%, respectively. In contrast, a network for CK+ and DISFA+ overall F1 score is 62% for both.

Our Experiment also shows that accuracy is dependent upon the number of AUs. We can improve accuracy by focusing on a few AUs with more training samples.

Finally, this detection of Facial Action Units and their intensity intern used to map them to corresponding basic and compound facial emotion with reasonable accuracy.

Reference

[1]. C. Darwin and P. Prodger, The expression of the emotions in man and animals. Oxford University Press, USA, 1998.

[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.

[3]. Li, Shan, and Weihong Deng. "Deep facial expression recognition: A survey." arXiv preprint arXiv:1804.08348 (2018).

[4]. H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," Image and Vision Computing, vol. 31, no. 2, pp. 120–136, 2013.

[5]. S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," Proceedings of the National Academy of Sciences, vol. 111, no. 15, pp. E1454–E1462, 2014.

[6]. Guo, Jianzhu, Zhen Lei, Jun Wan, Egils Avots, Noushin Hajarolasvadi, Boris Knyazev, Artem Kuharenko, et al. "Dominant and complementary emotion recognition from still images of faces." IEEE Access 6 (2018): 26391-26403.

[7]. Ekman, P., and Friesen, W. V. (1978). Facial Action Coding System: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press.

[8]. Xia, X.L., Xu, C. and Nan, B., 2017. Facial expression recognition based on tensorflow platform. In ITM Web of Conferences (Vol. 12, p. 01005). EDP Sciences.

[9]. Wang, Y., Li, Y., Song, Y. and Rong, X., 2019. Facial Expression Recognition Based on Auxiliary Models. Algorithms, 12(11), p.227.

[10]. Ko, Byoung. "A brief review of facial emotion recognition based on visual information." sensors 18, no. 2 (2018): 401.

[11]. Corneanu, C.A., Simón, M.O., Cohn, J.F. and Guerrero, S.E., 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE transactions on pattern analysis and machine intelligence, 38(8), pp.1548-1568.

[12]. K. Lekdioui, R. Messoussi, Y. Ruichek, Y. Chaabi et R. Touahni, Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier», Signal Processing: Image Communication, vol. 58, pp. 300312, 2017.

[13]. G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 915–928, 2007.

[14]. Pandey, R.K., Karmakar, S., Ramakrishnan, A.G. and Saha, N., 2019. Improving Facial Emotion Recognition Systems Using Gradient and Laplacian Images. arXiv preprint arXiv:1902.05411.

[15]. Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[16]. Chollet F . Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258

[17]. Benitez-Quiroz, C.F.; Srinivasan, R.; Martinez, A.M. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5562–5570.

[18]. S. Kim and H. Kim, "Deep Explanation Model for Facial Expression Recognition Through Facial Action Coding Unit," 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), 2019, pp. 1-4, DOI: 10.1109/BIGCOMP.2019.8679370.

[19]. Sánchez-Lozano, E., Tzimiropoulos, G., & Valstar, M. (2018). Joint action unit localisation and intensity estimation through heatmap regression. arXiv preprint arXiv:1805.03487.

[20]. 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," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101

[21]. Y. Song, D. McDuff, D. Vasisht, A. Kapoor, Exploiting sparsity and co-occurrence structure for action unit recognition, in:2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 1, 2015, pp. 1–8.

[22]. Z. Wang, Y. Li, S. Wang, Q. Ji, Capturing global semantic relationships for facial action unit recognition, in 2013 IEEE International Conference on Computer Vision, 2013, pp. 3304–3311.

[23]. S.-J. Wang, B. Lin, Y. Wang, T. Yi, B. Zou, X. wen Lyu, Action units recognition based on deep spatial convolutional and multi-label residual network, Neurocomputing 359 (2019) 130 – 138.

[24]. K. Zhao, W. S. Chu, F. De la Torre, J. F. Cohn, H. Zhang, Joint patch and multi-label learning for facial action unit and holistic expression recognition, IEEE Transactions on Image Processing 25 (8) (2016) 3931–3946.

[25]. M. Pantic, M. Valstar, R. Rademaker and L. Maat, "Web-based database for facial expression analysis," 2005 IEEE International Conference on Multimedia and Expo, 2005, pp. 5 pp.-, DOI: 10.1109/ICME.2005.1521424.

[26]. M. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. Systems, Man, and Cybernetics, Cybernetics, Transactions on, 42(1):28–43, 2012.

[27]. B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011.

[28]. S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. TPAMI, (11):1940–1954, 2010.

[29]. R. Walecki, O. Rudovic, V. Pavlovic and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, pp. 1-8, DOI: 10.1109/FG.2015.7163137.

[30]. S Mohammad Mavadati et al. "DISFA: A spontaneous facial action intensity database". In: Transactions on Affective Computing (2013).

[31]. Zhiwen Shao et al. "Facial action unit detection using attention and relation learning". In: Transactions on Affective Computing (2019).

[32]. Guanbin Li et al. "Semantic Relationships Guided Representation Learning for Facial Action Unit Recognition". In: AAAI. 2019.

[33]. Xuesong Niu et al. "Local Relationship Learning With Person-Specific Shape Regularization for Facial Action Unit Detection". In: CVPR. 2019.

[34]. Wei Li et al. "Eac-net: Deep nets with enhancing and cropping for facial action unit detection". In: T-PAMI (2018).

[35]. C. Corneanu, M. Madadi, S. Escalera, Deep structure inference network for facial action unit recognition, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 298–313.

[36]. Geethu Miriam Jacob, Bjorn Stenger; "Facial Action Unit Detection With Transformers", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7680-7689

[37]. S. Mohammad Mavadati, P. Sanger, Mohammad H. Mahoor, "Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions", Computer Vision and Pattern Recognition Workshop, June 2016

[38]. K. Zhao, W. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3391–3399.

[39]. C. Ma, L. Chen, J. Yong, Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection, Neurocomputing 355 (2019) 35–47. doi:10.1016/j.neucom.2019.03.082

[40]. Z. Shao, Z. Liu, J. Cai, L. Ma, JAa-net: Joint facial action unit detection and face alignment via adaptive attention (2020).

[41]. J. Egede, M. Valstar et B. Martinez, Fusing deep learned and handcrafted features of appearance, shape, and dynamics for automatic pain estimation, 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017.

[42]. M.-T. Yang, Y.-J. Cheng et Y.-C. Shih, Facial expression recognition for learning status analysis, International Conference on Human-Computer Interaction, 2011.

[43]. K. Slimani, R. Messoussi, S. Bourekkadi et S. Khoulji, An intelligent system solution for improving the distance collaborative work, Intelligent Systems and Computer Vision (ISCV), 2017.

[44]. Mourão et J. Magalhães, Competitive affective gaming: winning with a smile, Proceedings of the 21st ACM international conference on Multimedia, 2013.

[45]. L. D. Riek et P. Robinson, Using robots to help people habituate to visible disabilities IEEE International Conference on Rehabilitation Robotics (ICORR), 2011.

[46]. S. Bourekkadi, S. khoulji, K. Slimani, R. Messoussi et M. L. Kerkeb, The design of a psychotherapy remote intelligent system, Journal of Theoretical & Applied Information Technology, vol. 93, no 1, 2016.

[47]. M. S. Bartlett, G. Littlewort, I. Fasel et J. R. Movellan, Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. Conference on Computer Vision and Pattern Recognition Workshop, CVPRW'03, 2003.