FACS

Abstract:

This paper provides a detailed comparison of our approach to **Action Unit (AU) recognition** with existing state-of-the-art methods. By employing advanced deep learning architectures like Vision Transformers and Swin Transformers, we aim to enhance both the accuracy and efficiency of AU detection in practical, real-world applications. In this review, we examine various aspects of AU recognition, such as feature extraction methods, model architectures, and performance evaluation metrics, in comparison to the current body of research. We conducted a thorough analysis of literature from well-established sources, including **arXiv**, **IEEE**, and **ResearchGate**, to compile the most relevant and up-to-date studies on the subject.

This comparative study not only highlights the strengths and limitations of the existing techniques but also identifies gaps in the current research, offering insights into how AU recognition systems can be further improved. Our results emphasize the critical role of innovative feature extraction strategies and the importance of optimizing models to boost the accuracy of AU detection. By exploring these areas, we aim to contribute to the ongoing development of more robust and scalable AU recognition systems.

Keywords: Facial Action Unit Recognition, Affective Computing, Deep Learning, Vision Transformers, Swin Transformers, Feature Extraction, Emotion Detection, Human-Computer Interaction, arXiv, Google Scholar, ResearchGate, Comparative Study, Machine Learning

Paper Structure

1.Introduction

Facial expressions play a crucial role in human communication, conveying emotions and intentions non-verbally. One of the most systematic ways to decode and analyse these expressions is through the **Facial Action Coding System** (FACS), a comprehensive framework developed by Paul Ekman and Wallace V. Friesen in 1978. FACS is based on identifying specific muscle movements, referred to as Action Units (AUs), which correspond to distinct facial muscle activations. Unlike traditional emotion recognition approaches, which categorize expressions into discrete emotions like happiness, sadness, or anger, FACS provides a finer level of granularity by focusing on the fundamental building blocks of expressions. By combining multiple AUs, FACS can represent a wide range of complex emotional states and nuanced expressions.**[6]**

The ability to automatically detect and classify AUs from facial images has important applications across various domains, such as psychology, human-computer interaction, healthcare, and security. In healthcare, for instance, AU recognition can assist in diagnosing disorders like **autism** or **depression** by analysing facial expressions of patients over time. In human-computer interaction, AU detection allows systems to become more emotionally intelligent, enabling more natural and effective communication between humans and machines.

Over the years, AU recognition has evolved significantly, especially with the advancement of machine learning and deep learning techniques. Traditional approaches for AU recognition relied heavily on hand-crafted features such as local binary patterns (LBP), Gabor filters, and histogram of oriented gradients (HOG)**[7]**. These methods were limited in their capacity to generalize across diverse conditions, such as varying lighting, occlusion, or head poses. While they achieved moderate success, they often struggled to maintain high accuracy in real-world scenarios.

With the rise of deep learning, especially convolutional neural networks (CNNs), the field of AU recognition experienced a paradigm shift. **CNN**-based models are capable of learning hierarchical feature representations directly from raw image data, bypassing the need for manual feature engineering. These models have significantly improved performance in recognizing AUs under more complex and variable conditions. However, despite their success, CNNs face limitations when it comes to capturing long-range dependencies across facial regions and global spatial relationships.

Recently, the introduction of **Vision Transformers** (ViT) and Swin Transformers has brought new possibilities to the field of AU recognition. These transformer-based models excel at capturing both local and global features by employing **self-attention** mechanisms, which allow the model to weigh the importance of different facial regions when identifying AUs. Vision Transformers, in particular, have been found to perform well in tasks that require a comprehensive understanding of spatial relationships, making them well-suited for AU detection. **Swin Transformers**, which introduce a hierarchical structure with shifted windows, further refine the process by improving computational efficiency and scalability.

In this review, we compare various deep learning architectures, including CNNs, ViTs, and Swin Transformers, in terms of their effectiveness in AU detection. Additionally, we examine the role of feature extraction techniques and how innovations in this area, such as attention mechanisms and **multi-scale feature fusion**, have contributed to improved performance. Moreover, we analyze different evaluation metrics used in AU recognition research, such as the F1 score and accuracy and explore how current methods fare in terms of these benchmarks.

The ability to accurately and efficiently detect AUs has profound implications across multiple disciplines. In the field of mental health, AU recognition can aid in detecting subtle expressions that may indicate underlying emotional states or **psychological conditions[8]**. For instance, patients with depression or anxiety often exhibit micro-expressions that are hard to perceive with the naked eye but can be captured through automated AU detection systems. This makes AU recognition a valuable tool in therapy and diagnostics, allowing healthcare providers to monitor patient progress or assess emotional well-being more objectively.

In the context of **human-computer interaction**, AU recognition enhances the development of emotionally aware systems. Applications range from virtual assistants that can respond empathetically to user emotions, to video games that adapt dynamically based on a player's facial expressions. Additionally, AU recognition plays a critical role in **affective computing[9]**, where machines are designed to interpret and respond to human emotions, making interactions more seamless and intuitive.

2.Literature search and strategy

2.1 search strategy

We have gone through papers from various research sites using search keywords and have filtered papers which are too old and also the papers which have used complicated methods(GNN) for action unit detection(action unit identification and action unit intensity identification). We found most papers through the keywords ViT, affective computing , Swin transformer and emotion detection and we have compared our work with the work done on different datasets mentioned in the papers which we will be looking at in this paper.

2.2. Inclusion/Exclusion criteria

We have considered the papers which are published in scientific journals and conference proceedings. We have taken 120 papers for analysis and filtered out 70 papers which have same implementation or the duplicates and also removed the Graph Neural Network implementation to detect facial action units , we have ViT , DNN,CNN and other traditional feature extraction techniques implemented papers after screening the selected research papers.

1. **Identification**
Records identified from **scientific journals** and **conference proceedings** (n=60)

2. **First Screening**

1. Papers using **Vision Transformers (ViT)**
2. **Deep Neural Networks (DNN)** used for Facial Action Unit (AU) detection
3. **Convolutional Neural Networks (CNNs)** for AU detection
4. Papers using **other feature extraction techniques**

- **Records excluded**
  Papers with Graph neural network implementation are removed as they are complex in implementation and understanding.(20 papers excluded).

3. **Included**
Studies included in the final review
(n=40)

The research primarily focuses on Facial Action Unit (AU) Detection, Emotion Recognition, and the use of advanced machine learning techniques, particularly transformer architectures.

Walking through some of the famous research papers in this field ->

<u>AUFormer: Vision Transformers are Parameter-Efficient Facial Action Unit Detectors**</u>**

introduces a novel approach to Facial Action Unit (AU) detection using Vision Transformers (ViTs) as a more parameter-efficient alternative to traditional convolutional neural networks (CNNs). AUFormer**[1]** addresses the challenges of **overfitting** in AU detection, which often arise due to the **limited availability** of AU-annotated datasets and the need for models with fewer learnable parameters.

AUFormer incorporates a **Mixture-of-Knowledge Expert (MoKE)** collaboration mechanism, where each expert is specialized for a specific AU. This method allows for the integration of personalized multi-scale and correlation knowledge, and these MoKEs collaborate to **inject aggregated information** into the **frozen** Vision Transformer backbone. This approach makes the model highly efficient, reducing the need for large parameter sets while maintaining or surpassing the performance of existing methods.

Additionally, AUFormer introduces a novel **Margin-truncated Difficulty-aware Weighted Asymmetric Loss (MDWA-Loss)**, designed to handle the imbalanced nature of AU detection. This loss function encourages the model to focus more on activated AUs, accounting for the difficulty in detecting unactivated ones and ignoring potentially mislabeled samples. The model performs well across various tasks, including within-domain and cross-domain detection, and exhibits robust generalization without relying on external data.

<u>Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment**</u>**

introduces an innovative end-to-end deep learning framework that simultaneously addresses facial action unit (AU) detection and face alignment, integrates AU detection and face alignment into a single, cohesive model.

A novel adaptive attention learning module[2] is introduced to refine attention maps for each AU. This mechanism allows the model to focus on relevant facial regions, enhancing the precision of local feature extraction critical for AU detection

This joint approach leverages the interdependence between the two tasks, enabling the model to benefit from shared features and mutual reinforcement. The framework employs multi-scale shared feature learning, capturing both fine-grained and high-level information. This design ensures that features pertinent to face alignment and AU detection are effectively extracted and utilized.

Extensive experiments conducted on benchmark datasets such as **BP4D** and **DISFA** demonstrate that the proposed framework significantly outperforms existing state-of-the-art methods in AU detection.

Deep Region and Multi-label Learning for Facial Action Unit Detection

innovative approach to facial action unit (AU) detection by combining region learning (RL) and multi-label learning (ML) techniques. RL focuses on identifying specific facial regions where AUs are most active, enhancing the specificity of detection, while ML leverages the strong correlations between different AUs to improve detection accuracy.

A key feature of the **DRML**[3] model is its novel "region layer," which uses feed-forward functions to highlight crucial facial regions, thus capturing the structural information of the face more effectively. This layer operates as an intermediary between locally connected layers (confined to individual pixels) and traditional convolution layers (which share kernels across the image). By jointly tackling the RL and ML tasks in a unified deep network, DRML enhances interaction between these two aspects, improving detection performance.

Experiments using the BP4D and DISFA datasets demonstrate that the DRML model achieves state-of-the-art performance, particularly in terms of F1-score and AUC, surpassing alternative methods across both datasets. The end-to-end trainable nature of the network ensures that it learns robust representations even in the presence of local variations within facial regions.

FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features

introduces a novel approach aimed at improving the generalization capabilities of AU (Action Unit) detection across different datasets. The key innovation of FG-Net**[4]** is its use of **pyramidal CNN interpreters** and features extracted from a **pre-trained StyleGAN2 model.**

the model addresses common challenges such as **overfitting** and poor cross-corpus performance that plague traditional AU detection methods. The system demonstrates high efficiency and robustness, particularly in **cross-domain** settings, where models typically struggle. FG-Net has been tested on the widely-used **DISFA** and **BP4D** datasets, showing superior generalization while maintaining competitive performance within individual datasets. It also proves to be **data-efficient**, requiring as few as 1000 training samples to achieve notable results.

FG-Net serves as a leading example of how generative models and pyramid networks can boost performance, both within and across datasets, providing a promising direction for future research.

** Emotion Recognition Using Transformers with Masked Learning**

emotion recognition that leverages Vision Transformers **[**5**]**(ViT) for more effective analysis of emotional states. The research focuses on estimating valence-arousal (VA), recognizing facial expressions, and detecting action units (AUs) that represent facial muscle movements, key components in understanding emotions.

The authors introduce a unique masked learning strategy where frames of a video are randomly masked during training, encouraging the model to learn more robust temporal and spatial features. This approach enhances the performance of Transformers compared to traditional methods such as CNNs or LSTMs. Moreover, the paper **adopts focal loss(This is the paper which we referred to use focal loss in our work)** to address the challenge of class imbalance often found in datasets related to emotion recognition, particularly in real-world, in-the-wild settings.

The experimental results, particularly with the Aff-Wild2 dataset, show improvements in VA estimation and AU detection, making this work highly applicable to affective computing and real-world emotion detection scenarios

3.Facial action unit fundamentals

In the context of mental health, facial Action Unit (AU) recognition plays a pivotal role in understanding and detecting stress and emotional states, offering potential breakthroughs in stress detection applications. Stress can manifest through subtle facial changes—muscle movements that might not be visible to the naked eye but can be detected through the precise recognition of AUs.

For a stress-detecting application aimed at assisting psychologists, this technology allows for an objective, real-time analysis of a patient's emotional state. It works by recognizing facial muscle movements that are linked to stress-related expressions. For instance, stress might be identified through combinations of AUs like brow raising (AU1), eye narrowing (AU7), or lip tightening (AU23), which can correlate with emotional strain or anxiety.

By integrating AU recognition into a mental health tool, psychologists can receive continuous feedback on a patient’s stress levels, even during regular therapy sessions or day-to-day activities. This could help therapists better understand the emotional triggers and responses of their patients and customize interventions accordingly. Additionally, the application could offer remote monitoring capabilities, where the system detects facial cues of stress and alerts caregivers or clinicians in real-time, providing valuable insights into the patient's mental state between therapy sessions.

Such a system offers immense potential in preventing burnout, anxiety, and chronic stress by recognizing early signs of emotional distress. It can assist psychologists in designing better treatment plans based on consistent emotional monitoring and ensuring that patients receive the timely support they need.

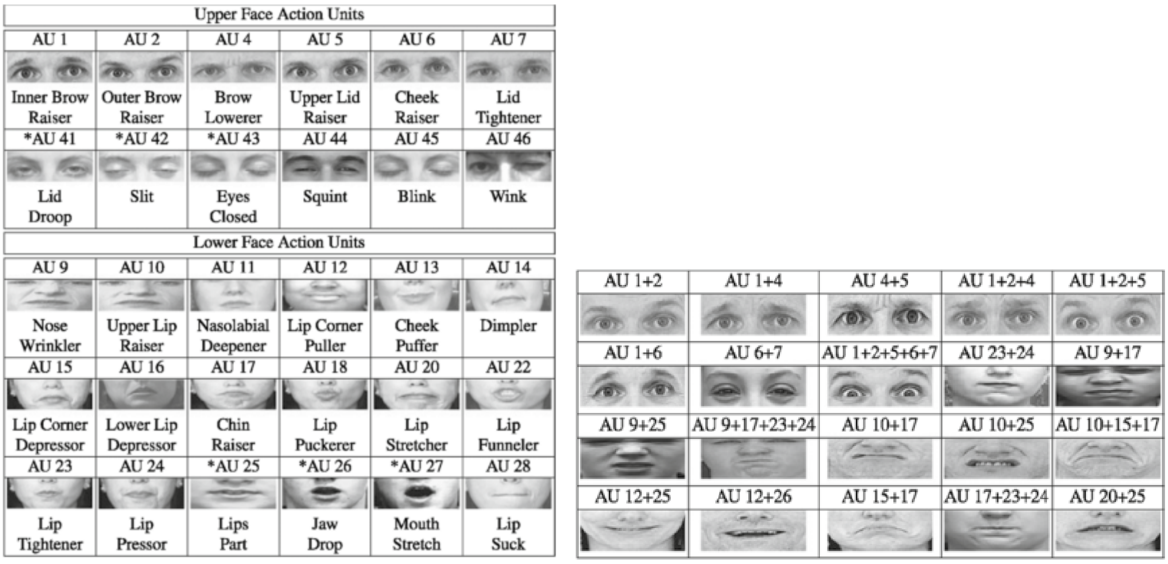


Figure 1.Facial action units(individual and combination)

3.1 Definition and Representation of AUs

The intensity of Facial Action Units (AUs) refers to the degree to which a specific facial movement is executed. Accurately representing AU intensity is crucial for understanding the emotional state of an individual, as it provides insights into the strength of a particular expression. This representation can be used in applications ranging from psychology to affective computing and human-computer interaction.

1. Measurement of AU Intensity

AU intensity is typically measured on a scale from 0 to 5, where:

- **0** indicates no visible movement,
- **1** represents a trace amount of movement,
- **2** is a slight movement,
- **3** is moderate,
- **4** is strong, and
- **5** indicates maximum intensity.

The intensity of AUs can significantly impact the interpretation of emotions. For instance, a smile (involving AU12) that registers as a strong intensity (AU12 = 4 or 5) may convey happiness or joy, whereas a slight smile (AU12 = 1 or 2) may imply politeness or discomfort. Therefore, systems that measure and analyze AU intensity are better equipped to classify emotions accurately.

In the context of mental health, the representation of AU intensity can be particularly beneficial for stress detection applications. By analyzing the intensity of specific AUs, psychologists can gain insights into a patient's emotional state and stress levels. For instance, increased intensity in AUs associated with negative emotions could indicate rising stress or anxiety, allowing for timely intervention and support

3.2 Feature extraction techniques

Histogram of Oriented Gradients (HOG) is a feature extraction technique that focuses on the distribution of gradient orientations in localized sections of an image. It effectively highlights the edges and contours of facial features, making it particularly useful for recognizing facial expressions tied to specific action units (AUs). HOG's**[**10**]** robustness against variations in lighting and pose helps maintain detection accuracy across different environments. By emphasizing spatial information and maintaining orientation invariance, HOG is instrumental in identifying the subtle changes in facial geometry that characterize expressions of emotions like happiness, anger, or surprise .

Local Binary Patterns (LBP)

Local Binary Patterns (LBP)**[**11**]** is a texture descriptor that encodes local patterns by comparing each pixel with its surrounding neighbors, creating a binary code that reflects variations in texture. This technique is particularly valuable in facial action unit detection, as it captures minute changes in skin texture resulting from muscle movements during expressions. LBP's computational efficiency allows for real-time applications, making it suitable for systems requiring fast response times, such as video analysis and surveillance. Its strength lies in its ability to maintain invariance to monotonic grayscale changes, thus providing reliable features for identifying different facial expressions associated with various AUs .

Gabor Filters

Gabor filters are a set of linear filters used to analyze spatial frequency content and texture in images. By applying these filters at different orientations and scales, Gabor filters**[**12**]** can capture localized features that correspond to specific facial movements. This capability makes them particularly effective for AU detection, as they can identify the fine details of muscle contractions around the eyes and mouth, which are crucial for expressions like surprise or fear. Gabor filters are adept at handling variations in illumination and are sensitive to edge orientations, enhancing their efficacy in distinguishing between subtle changes in facial expressions .

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are deep learning models that excel at automatically learning hierarchical feature representations from raw pixel data. In the context of Facial Action Unit detection, CNNs**[**13**]** can effectively capture complex patterns and variations associated with facial expressions without the need for manual feature engineering. The layered architecture of CNNs allows them to learn both local and global features, enabling them to discern subtle differences in facial movements that signify different AUs. Their robustness to variations in lighting, occlusions, and facial orientations makes CNNs a popular choice for real-time applications in emotion recognition and facial analysis .

Vision Transformers (ViTs)

Vision Transformers (ViTs)**[**14**]** represent a novel approach to image analysis by treating images as sequences of patches, allowing for the capture of long-range dependencies and global context within images. This architecture is particularly beneficial for Facial Action Unit detection, as it enables the model to consider the relationship between different facial regions, which is crucial for understanding complex expressions. ViTs have shown promising results, often surpassing traditional CNNs in performance, especially when trained on large datasets. Their ability to leverage attention mechanisms enhances the model's capacity to focus on relevant features that signify distinct AUs, contributing to more accurate emotion recognition .

Feature extraction techniques adapted/Exploited in **our work**:

1.CNN – we have used MobileNet,InceptionV3,Efficientnet B2 pretrained models as CNN feature extraction techniques where we have removed the bottleneck dense layers and added our layers to finetune the model as per our task.

2.ViT – We have used pretrained vision transformer trained on ImageNet dataset from google(google-vit-patch16) and Meta(dino) as feature extractors producing 1024 and 768 features respectively which have produced good results on the test dataset(part of DISFA).

3.LBP,HOG and Gabor Filters

Extracting all three features and concatenating them to train a neural network to detect au's from an image but due to similarity in features in the images due to lack of availability of diverse subjects in the dataset, the model is underfitting and could not detect Action unit with a good reasonable accuracy.

4.Configural Features

(We have found this in a research paper from **IIT Hyderabad** "Configural Representation of Facial Action Units for Spontaneous Facial Expression Recognition in the Wild")**[**15**]**.

These are features closely related to facial landmarks which serve as reference points allowing us to analyze spatial relationship between different features such as distance between landmarks as a features , we can also use to calibrate our predictions from models built through other techniques.

4.Datasets and benchmarking

4.1 Publicly Available AU Datasets

Dataset used in our work

DISFA(Denver intensity of Spontaneous Facial Action)**[**

videos of 27 individuals with 4845 frames each annotated with 12 action units

non-posed facial expression database**[**16**]** for those who are interested in developing computer algorithms for automatic action unit detection and their 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). The intensity of AU's (0-5 scale) for all video frames were manually scored by two human FACS experts. The database also includes 66 facial landmark points of each image in the database. Twenty-seven young adults were video-recorded by a stereo camera while they viewed video clips intended to elicit spontaneous emotion expression. Participants viewed a 4-minute video clip (242 seconds in length) intended to elicit spontaneous AUs in response to videos intended to elicit a range of facial expressions of emotion. The clip consisted of 9 segments taken mostly from YouTube. Further information about the video clip is provided in the Appendix. While viewing the video, participants sat in a comfortable chair positioned in front of a video display and stereo cameras. They were alone with no one else present. Their facial behavior was imaged using a high-resolution (1024 × 768 pixels) BumbleBee Point Grey stereo-vision system at **20** fps under uniform illumination. For each participant, **4845** video frames were recorded.The imaging system is depicted in Figure 3. AU intensity was coded for each video frame on a 0 (not present) to 5 (maximum intensity) ordinal scale. For each AU, we report the number of events and the number of frames for each intensity level. Event refers to the continuous occurrence of an AU from its onset (start frame) to its offset (end frame) . FACS coding was performed by a single **FACS coder**.

Other Datasets on Facial action units

S.NO	Dataset	Subjects count	Size per subject	Frame level labelling
1	DISFA+	9	4845	yes

2	CK+	201	593	yes
3	AMFED	5268	242	yes
4	BP4D	41	328	yes
5	CASME	35	247	no
6	EmotioNet	NA	950000frames	yes

4.2 Evaluation metrics for AU recognition models

We have used F1-Score , Accuracy and PR-AUC to assess our models performance during training and testing on DISFA dataset.

F1-Score is the harmonic mean of Precision and Recall giving us a balance between both the values which are important for our analysis where False Positives and False negatives play a crucial analysis in stress detection.

False Positives ⬇ Precision ⬆

False Negatives ⬇ Recall ⬇

Choosing F1-score as a metric is due to the imbalance in the dataset where accuracy as a metric will be fail to gain true insights about the model, **PR-AUC** does the same thing by letting us know the area under Precision recall curve and choosing the threshold for binary classification of each AU which maximizes the F1-Score.

Confusion Matrix is also used to visualize the mis predictions being done by the model for each action unit

5.Comparative analysis

5.1 Our Work

The proposed approach for facial action unit recognition contains 4 major steps

(i) Face and landmark detection
(ii) Facial Image preprocessing ,
(iii) Facial action unit (FAUs) recognition.
(iv) Mapping Detected Action units to emotions(Micro and Macro).**[**17**]**

Our focus is more on step 3,4 and comparing other works and results to ours.

1.For face detection, we used the Dlib model **res10_300x300_ssd_iter_140000 [**11**]**, which is a pre-trained Single Shot Multibox Detector (SSD) model. This model operates on a resolution of 300x300 pixels and is designed to efficiently detect faces in images.

2.As a part of facial image preprocessing To standardize the input for our model and optimize computational efficiency, all images were resized to a fixed resolution of 224x224 pixels with three color channels (RGB). This resizing ensures consistency in input dimensions, allowing the model to process images uniformly. After resizing, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) **[**9**]** to each image.

CLAHE is an advanced histogram equalization technique that enhances the contrast of images, particularly useful for improving the visibility of features in facial regions. In our study, we experimented with various clip limits and grid sizes to optimize the contrast enhancement. The clip limit parameter controls the contrast enhancement degree by limiting the maximum slope of the cumulative distribution function, while the grid size determines the size of the contextual regions used for histogram equalization. By adjusting these parameters, we tailored the preprocessing step to maximize the visibility of facial landmarks and action units, ultimately improving the model's feature extraction and classification performance.Furthermore, to ensure that the preprocessing pipeline was optimized for feature extraction, we conducted a series of experiments to fine-tune the **CLAHE** parameters, such as **clip limit** and **grid size**, which are crucial for maximizing the visibility of subtle facial features.

We have even reduced the count of samples which have no action unit in them as the dataset is highly skewed towards it. We came up by eliminating 20k such frames inorder to balance it but it is still skewed which is a drawback of disfa dataset.

3.Action unit detection

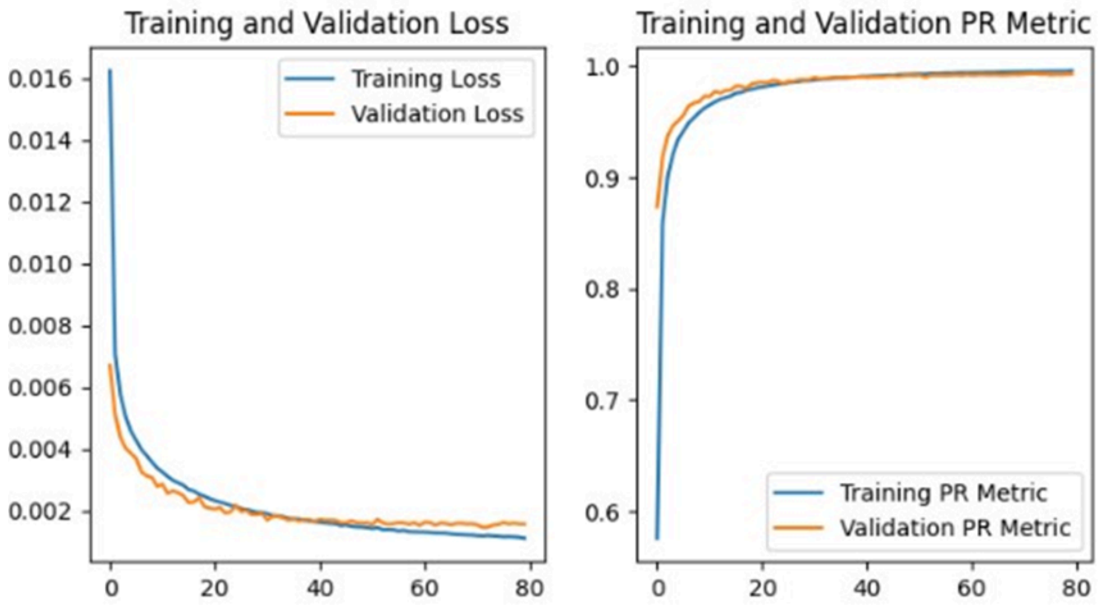
We have divided our task into 2 parts (action unit detection and action unit intensity detection) currently working on detecting presence of action unit .

After trying various models with CNN-based extraction, traditional Feature extraction and Ensemble methods we have come up with 3 models XGBOOST,meta(dino) based ViT and InceptionV3 models which have produced good numbers on DISFA dataset when trained on 20 samples of DISFA and tested on 7 samples of DISFA.

Training Process:

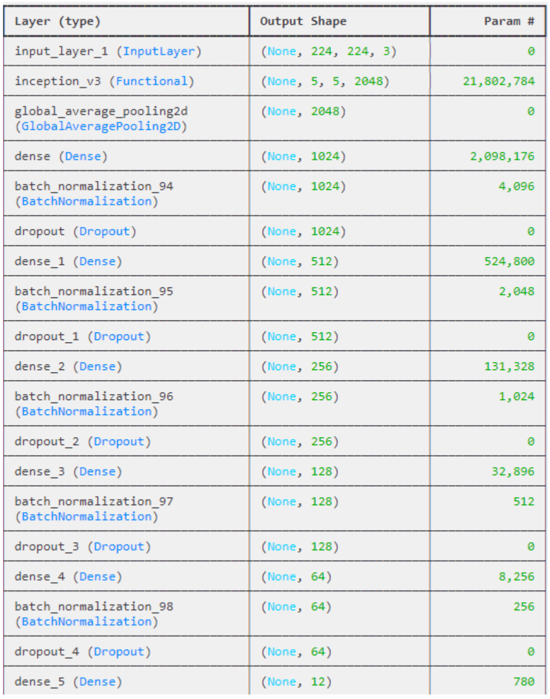
Finetuning **Vision Transformer** for Au detection**[**18**]**

- Trained on Kaggle GPU P100
- Framework used : Tensorflow
- Optimizer : Adam
- Batch size : 128
- Loss Function : Focal Loss (to handle class imbalance)
- Epochs : 80
- Train, validation, Test split : 52:13:35 percent.



We have used pretrained Vision transformers and fine tuned them for au-detection task and also some other SOTA Models In image classification

Pretrained InceptionV3 trained on ImageNet dataset



Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 224, 224, 3)	0
inception_v3 (Functional)	(None, 5, 5, 2048)	21,802,784
global_average_pooling2d (GlobalAveragePooling2D)	(None, 2048)	0
dense (Dense)	(None, 1024)	2,098,176
batch_normalization_94 (BatchNormalization)	(None, 1024)	4,096
dropout (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 512)	524,800
batch_normalization_95 (BatchNormalization)	(None, 512)	2,048
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131,328
batch_normalization_96 (BatchNormalization)	(None, 256)	1,024
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 128)	32,896
batch_normalization_97 (BatchNormalization)	(None, 128)	512
dropout_3 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 64)	8,256
batch_normalization_98 (BatchNormalization)	(None, 64)	256
dropout_4 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 12)	780

- We are using Global average pooling to reduce the parameters or the features from pre trained model to reduce overfitting and also by taking the average global spatial information of feature maps retain.

- Then multiple Fully connected layers with dropout (to avoid overfitting) to map the features to the output labels with 12 neurons in output layer.

Training Process:

- Trained on Kaggle GPU P100
- Framework used : Tensorflow
- Optimizer : Adam
- Batch size : 32
- Loss Function : Focal Loss (to handle class imbalance)
- Callbacks : Early Stopping (Monitoring validation loss)
- Epochs : 50
- Train, validation split : 80:20 (65420 images before sampling,45116 after sampling).
- Sampling in train data : as there were many frames where no action unit is present(24.3k) out of 65k , sampled 4k frames randomly to denote no action unit class.

Ensemble techniques with XGBoost

The implementation of a multi-label classification model using XGBoost with GPU acceleration presents a robust approach for facial action unit (AU) detection. XGBoost, recognized for its efficiency and high predictive power, employs gradient boosting techniques that iteratively improve model performance by learning from the errors of previous iterations. By wrapping the XGBoost model within a MultiOutputClassifier, the framework accommodates multi-label outputs, enabling the simultaneous prediction of various AUs

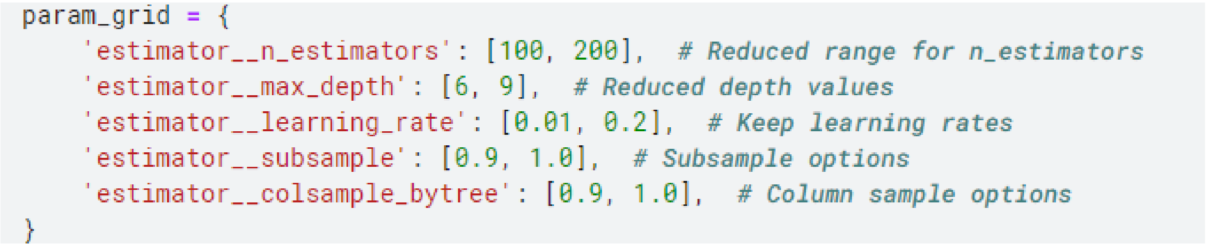
- Trained on Kaggle GPU P100

- Framework used : Scikit-Learn

- Loss Function : Binary-Logistic

- Fits : 2 Fold CV with 32 fits each

- Params :

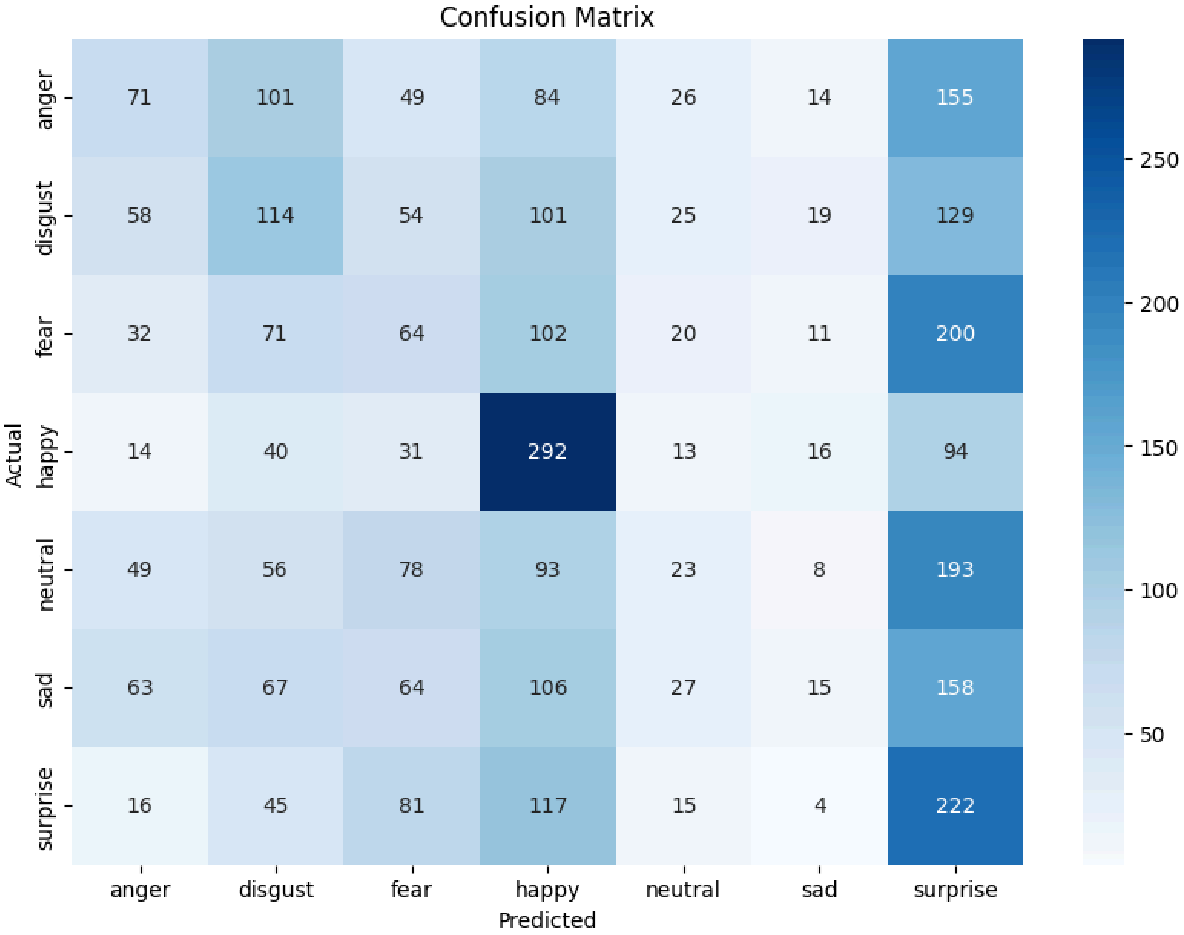


```
param_grid = {
    'estimator__n_estimators': [100, 200],  # Reduced range for n_estimators
    'estimator__max_depth': [6, 9],  # Reduced depth values
    'estimator__learning_rate': [0.01, 0.2],  # Keep learning rates
    'estimator__subsample': [0.9, 1.0],  # Subsample options
    'estimator__colsample_bytree': [0.9, 1.0],  # Column sample options
}
```

Figure 2.param grid for training and cross validation

- Train, Test split : 74:26 percent.

4.But when we have tried to map these detected action units to emotions for images taken from AffectNet dataset which is a famous facial expression recognition dataset , models numbers are not too good (28% accuracy and 0.2 F1 score) indicating we are overfitting It.



20

Figure 4.XGBmodel when mapped action units to emotions

To Map Detected action units to Emotions we have used traditional facs ruling mentioned

by paul ekman group

We are classifying the action units into 7 Major emotions

```
emotions = {

        'anger': ['au4', 'au5', 'au17'],

        'Sad': ['au1', 'au4', 'au15'],

        'happy': ['au12', 'au6'],
        'Fear': ['au1', 'au2', 'au5', 'au20'],

        'Disgust': ['au9', 'au15'],

        'Surprise': ['au1', 'au2', 'au5', 'au26']

    }
```

And no action unit detected indicates neutral , incase of class of au’s or missing of a particular au we assumed it to be present and went for maximum percentage match to get emotion out of it

After all these trials we have chosen XGB/InceptionV3 as final choices for the model as it has number very close to **ViT** but speed plays a crucial role here as Vit is a heavy model it is taking comparatively much longer than XGB,**InceptionV3** as it is developed by **Google** which has SOTA performance at the ImageNet Recognition Challenge. We can choose GPU supported XGB model once it outperforms other models as it is the fastest model right now.

5.2 Comparing our work done with benchmarks models

Our Models on DISFA

S.No	Model	Dataset	No of au’s	Avg F1 score
1	**XGB(OUR)**	**DISFA**	**12**	**0.4**
2	**VIT(OUR)**	**DISFA**	**12**	**0.42**

Different models mentioned in paper Towards End-to-End Explainable Facial Action Unit Recognition via Vision-Language Joint Learning [19]**(ours is only action unit detection task and model is overfitted refer fig4).**

Method	1	2	4	6	9	12	25	26	Avg.
XGB model(Ours)	86	89	93	93	93	94	95	89	91.5

UGN-B (AAAI2021)	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	59.40
JAA-Net (ECCV2019)	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.04
FAU-Trans (CVPR2021)	46.1	48.5	72.8	56.7	50.0	72.1	90.8	65.0	62.75
ME-GraphAU (IJCAI2021)	54.6	72.9	54.0	55.7	76.7	91.1	53.0	62.1	65.01
KDSRL (CVPR2022)	67.5	52.7	67.5	76.1	91.3	57.7	62.4	64.2	67.42
KS (ICCV2023)	53.8	59.9	69.2	54.2	75.8	92.2	46.8	64.8	64.59
AAR (TIP2023)	62.4	39.0	48.8	76.1	91.3	70.6	64.2	63.9	64.54
SMA-ViT (TAC2023)	51.2	49.3	64.7	48.3	87.6	85.1	61.2	62.2	63.70

performance of models on EmotioNet Dataset as a part of EmotioNet challenge

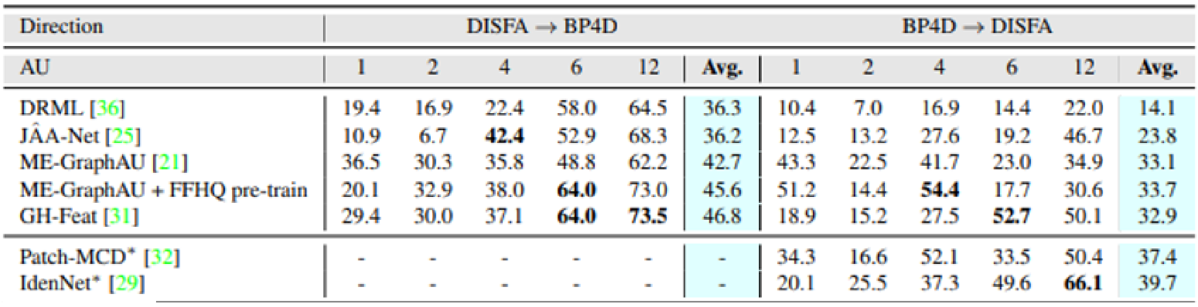
Year	References	F.5 score	F2 score	F1 score	Final score	Accuracy
2017	I2R-CCNU-NTU-2	0.64	0.643	0.641	0.729	0.82
	JHU	0.638	0.635	0.632	0.702	0.763
	I2R-CCNU-NTU-1	0.635	0.625	0.626	0.699	0.782
	I2R-CCNU-NTU-3	0.627	0.62	0.62	0.694	0.775
2018	PingAn-Gamma Lab				0.7553	0.9446
	VisionLabs				0.6718	0.9207
	MIT				0.6711	0.9298
	University of Washington				0.6300	0.8869
	PingAn-Tech				0.6221	0.8694

	University of Denver				0.5436	0.8576
	*KSDA	0.621	0.611	0.619	0.710	0.807
	*AlexNet	0.353	0.446	0.266	0.515	0.763

Performance of various models on 2017 FERA sub challenge for action unit detection

Year	Database and AUs	Registration	Representation	Classifier	
2017	BP4D+(AU1, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23)	Morphology operations & Binary segmentation	VGG-Face	0.574 0.778	
2017		SeetaFace Engine **[**215**]**	CNN & BLSTM-RNN	0.507 0.735	
2017		Faster R-CNN **[**144**]**	AUMPNet	0.506 -	
2017		SeetaFace **[**163**]**	CNN, Random Forest	0.498 0.694	
2017		*Valstar et al. **[**198**]**	Cascaded Continuous Regression **[**175, 176**]**	CRF	0.452 0.561

Cross-dataset validation on common Action unit performance by famous models



Direction	DISFA → BP4D						BP4D → DISFA					
AU	1	2	4	6	12	**Avg.**	1	2	4	6	12	**Avg.**
DRML [36]	19.4	16.9	22.4	58.0	64.5	36.3	10.4	7.0	16.9	14.4	22.0	14.1
JÂA-Net [25]	10.9	6.7	**42.4**	52.9	68.3	36.2	12.5	13.2	27.6	19.2	46.7	23.8
ME-GraphAU [21]	36.5	30.3	35.8	48.8	62.2	42.7	43.3	22.5	41.7	23.0	34.9	33.1
ME-GraphAU + FFHQ pre-train	20.1	32.9	38.0	**64.0**	73.0	45.6	51.2	14.4	**54.4**	17.7	30.6	33.7
GH-Feat [31]	29.4	30.0	37.1	**64.0**	**73.5**	46.8	18.9	15.2	27.5	**52.7**	50.1	32.9
Patch-MCD* [32]	-	-	-	-	-	-	34.3	16.6	52.1	33.5	50.4	37.4
IdenNet* [29]	-	-	-	-	-	-	20.1	25.5	37.3	49.6	**66.1**	39.7

Figure 8.Metrics from FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features

5.3 Limitation of current systems

Facial Action Coding System (FACS) is a powerful tool for analyzing facial expressions by identifying Action Units (AUs), which correspond to specific muscle movements. However, manual AU annotation is **time-consuming** and **labor-intensive**, requiring significant training and effort. This process can take over 100 hours for coders to achieve basic competency and even more time to annotate images or videos. Moreover, there is inconsistency in labeling, as inter-annotator reliability can be challenging to maintain.**[**20**]**

Another major obstacle to FACS is that frequent head movements and subtle facial deformations complicate labeling. Automatic annotation systems promise faster, more consistent AU labeling while improving accuracy and temporal resolution for real-world applications. While automatic AU occurrence detection is well-researched, AU intensity estimation remains a less explored area, even though it offers critical insights into the complexity of facial behaviors, such as distinguishing between genuine and posed emotions.

Moreover, a significant limitation in AU recognition research is the lack of publicly available, diverse datasets with annotated AUs. This scarcity hinders model training and generalization, especially when trying to capture spontaneous, real-world facial behaviors. Most available datasets include limited action units and do not cover diverse populations or settings, further restricting the development of robust AU recognition systems. Addressing this limitation requires expanding dataset availability and incorporating more nuanced AU combinations in research.

we have very few **publicly available** datasets for research purpose whereas others come with a fee which have diverse subjects and more action units annotated. Lack of **resources** is also an issue when building a DL/AI model as the framesize and the training data increase the need for GPU to train the model increases which is not free in most cases.

6. Conclusion

6.1 Summary of Key Findings

This review provided an in-depth analysis of the advancements in Facial Action Unit (AU) recognition, focusing particularly on the progression from traditional feature-based approaches, such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), to more modern deep learning architectures. The key finding was that models utilizing Vision Transformers (ViTs) and Swin Transformers significantly outperform earlier Convolutional Neural Networks (CNNs) in detecting and classifying AUs. These transformer-based models are particularly effective because of their ability to capture both local and global features across facial regions, allowing for a more accurate and nuanced interpretation of facial expressions.

We also highlighted that attention mechanisms and multi-scale feature fusion contribute substantially to improving AU recognition performance. This comparative analysis showed that, while traditional approaches such as CNNs are still useful in certain contexts, transformer models offer superior capabilities, especially in complex real-world applications. Moreover, our results indicate that innovative loss functions like the Margin-truncated Difficulty-aware Weighted Asymmetric Loss (MDWA-Loss) help address class imbalance issues in AU datasets, further improving model accuracy and robustness.

6.2 Impact of AU Recognition in Various Domains

Facial Action Unit recognition has the potential to transform several domains due to its ability to interpret subtle facial expressions with high precision. In **psychology and mental health**, AU recognition plays a critical role in monitoring emotional states, stress levels, and behavioral patterns. For example, patients suffering from anxiety or depression often display micro-expressions that might be too subtle for the human eye to detect but can be captured by AU recognition systems. This technology enables therapists to assess emotional fluctuations more objectively, offering real-time insights into a patient's mental health and aiding in early diagnosis or tracking progress during treatment.

In **human-computer interaction (HCI)**, AU detection enhances the emotional intelligence of systems, allowing machines to interpret and respond to human emotions more naturally. This has significant implications for fields like virtual assistants, customer service, and video games, where emotionally adaptive responses can improve user experience. Systems equipped with AU recognition can adjust their responses based on real-time user emotions, making interactions feel more personal and engaging.

In **security and surveillance**, AU recognition provides a new layer of emotional and behavioral analysis. Detecting stress, anxiety, or nervousness through facial expressions can help identify suspicious or high-risk individuals in real-time, enhancing security measures in airports, public spaces, or sensitive facilities. Similarly, **healthcare** applications benefit from AU recognition for monitoring patients’ emotional well-being over time, particularly in remote settings where continuous observation is otherwise impractical.

Finally, in **education**, AU recognition can be used to gauge student engagement and emotional responses during virtual or in-person lessons, providing educators with feedback that allows for more personalized and effective teaching methods.

6.3 Final Thoughts and Future Work

Despite the progress made in AU recognition, several challenges remain that present opportunities for future research. One major limitation is the **lack of diverse, large-scale datasets** that include spontaneous and natural expressions across varied populations. Most current datasets are either too small or lack diversity in terms of ethnicity, age, and environmental conditions. The absence of such datasets limits the generalizability of AU recognition models in real-world scenarios. Future work should focus on creating or gaining access to more comprehensive datasets that cover a broader spectrum of facial expressions, demographics, and conditions (e.g., lighting, occlusion, head poses).

Another area that requires further exploration is the **estimation of AU intensity**. While detecting the presence of AUs is important, understanding the intensity of these AUs—whether a person is smiling slightly or broadly—provides deeper insight into the underlying emotions. However, intensity estimation remains under-researched, primarily due to the lack of annotated datasets that provide intensity scores. More robust models need to be developed to handle AU intensity prediction, which can then be applied in domains like mental health, where the subtlety of an expression might be critical for diagnosis.

Furthermore, **multi-modal approaches**—integrating facial expressions with other data sources like speech, physiological signals, or body language—are another promising direction for future work. Combining these data sources could lead to more accurate and context-aware emotion recognition systems. Additionally, **real-time performance** and **computational efficiency** need improvement, especially in large-scale applications where speed is critical, such as surveillance or interactive systems. Optimizing transformer-based models for faster inference without sacrificing accuracy will be crucial for deploying AU recognition in practical, real-world environments.

In conclusion, the development of Facial Action Unit recognition systems holds immense potential across a variety of domains. While current models show promising results, there is still a need for more diverse datasets, improved intensity estimation techniques, and the integration of multi-modal data for even more comprehensive and robust systems. With ongoing research and innovation, AU recognition will continue to evolve, providing deeper insights into human emotions and further bridging the gap between humans and machines in emotionally intelligent interactions.

7. References

1. AUFormer: Vision Transformers are Parameter-Efficient Facial Action Unit Detectors

Kaishen Yuan, Zitong Yu, Xin Liu, Weicheng Xie, Huanjing Yue, Jingyu Yang

2.Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment

Zhiwen Shao, Zhilei Liu, Jianfei Cai, Lizhuang Ma

3.K. Zhao, W. -S. Chu and H. Zhang, "Deep Region and Multi-label Learning for Facial Action Unit Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 3391-3399, doi: 10.1109/CVPR.2016.369.

4.FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features

Yufeng Yin, Di Chang, Guoxian Song, Shen Sang, Tiancheng Zhi, Jing Liu, Linjie Luo, Mohammad Soleymani

5.Emotion Recognition Using Transformers with Masked Learning

Seongjae Min, Junseok Yang, Sangjun Lim, Junyong Lee, Sangwon Lee, Sejoon Lim

6. J. Yang, F. Zhang, B. Chen and S. U. Khan, "Facial Expression Recognition Based on Facial Action Unit," *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, Alexandria, VA, USA, 2019, pp. 1-6, doi: 10.1109/IGSC48788.2019.8957163.

7. DISFA: A Spontaneous Facial Action Intensity Database S.Mohammad Mavadati, Student Member, IEEE, Mohammad H. Mahoor, Member, IEEE, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn

8. Towards Independent Stress Detection: A Dependent Model Using Facial Action Units

Carla Viegas; Shing-Hon Lau; Roy Maxion; Alexander Hauptmann

9. Automatically Recognizing Facial Indicators of Frustration: A Learning-centric Analysis

Joseph F. Grafsgaard; Joseph B. Wiggins; Kristy Elizabeth Boyer; Eric N. Wiebe; James C. Lester

10. A real-time robust facial expression recognition system using HOG features

Pranav Kumar; S L Happy; Aurobinda Routray

11. Local normal binary patterns for 3D facial action unit detection

Georgia Sandbach; Stefanos Zafeiriou; Maja Pantic

12. Recognizing facial actions using Gabor wavelets with neutral face average difference

J.J. Bazzo; M.V. Lamar

13. Leveraging Previous Facial Action Units Knowledge for Emotion Recognition on Faces

Pietro B. S. Masur; Willams Costa; Lucas S. Figueredo; Veronica Teichrie

14. Facial Action Unit Detection with ViT and Perceiver Using Landmark Patches

Duygu Cakir; Gorkem Yilmaz; Nafiz Arica

15. Perveen, Nazil & Mohan, Chalavadi. (2020). Configural Representation of Facial Action Units for Spontaneous Facial Expression Recognition in the Wild. 93-102. 10.5220/0009099700930102.

16. Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions

Mohammad Mavadati; Peyten Sanger; Mohammad H. Mahoor

17. A method to infer emotions from facial Action Units

Sudha Velusamy; Hariprasad Kannan; Balasubramanian Anand; Anshul Sharma; Bilva Navathe

18. Multi-Modal Learning for AU Detection Based on Multi-Head Fused Transformers

Xiang Zhang; Lijun Yin

19. Towards End-to-End Explainable Facial Action Unit Recognition via Vision-Language Joint Learning

Xuri Ge, Junchen Fu, Fuhai Chen, Shan An, Nicu Sebe, Joemon M. Jose

20. Namba, S.; Sato, W.; Osumi, M.; Shimokawa, K. Assessing Automated Facial Action Unit Detection Systems for Analyzing Cross-Domain Facial Expression Databases

21. DeepFN: Towards Generalizable Facial Action Unit Recognition with Deep Face Normalization Javier Hernandez, Daniel McDuff, Ognjen (Oggi) Rudovic, Alberto Fung, and Mary Czerwinski

22. Facial Expression Recognition Based on Facial Action Unit

Jiannan Yang; Fan Zhang; Bike Chen; Samee U. Khan

23. Enhanced Facial Expression Recognition Based on Facial Action Unit Intensity and Region Weiyang Chen; Anrui Wang

24. Attention Based Relation Network for Facial Action Units Recognition

Yao Wei; Haoxiang Wang; Mingze Sun; Jiawang Liu

25. Action unit reconstruction of occluded facial expression

Chung-Hsien Wu; Jen-Chun Lin; Wen-Li Wei

26. Facial action unit recognition using temporal templates

M. Valstar; I. Patras; M. Pantic

27. Probabilistic inference of facial action unit by Dynamic Bayesian Network for image sequence Yee Koon Loh; Shahrel A. Suandi

28. Self-supervised Facial Action Unit Detection with Region and Relation Learning

Juan Song, Zhilei Liu

29. Multi-Task Transformer with uncertainty modelling for Face Based Affective Computing Gauthier Tallec, Jules Bonnard, Arnaud Dapogny, Kévin Bailly

30. Transformer-based Multimodal Information Fusion for Facial Expression Analysis

Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, Yu Ding

31. P. Kumar, S. L. Happy and A. Routray, "A real-time robust facial expression recognition system using HOG features,"

32. S. M. Mavadati, M. H. Mahoor, K. Bartlett and P. Trinh, "Automatic detection of non-posed facial action units," 2012

33. P. Siritanawan and K. Kotani, "Facial action units detection by robust temporal features,"

34. S. C. Bakchy, M. J. Ferdous, A. H. Sathi, K. C. Ray, F. Imran and M. M. Ali, "Facial Expression Recognition based on Support Vector Machine using Gabor Wavelet Filter"

35. Facial Action Unit Recognition Based on Transfer Learning Shangfei Wang, Yanan Chang, Jiahe Wang University of Science and Technology of China, Hefei, Anhui, China

36. S. F. Wang, J. Xue and X. F. Wang, "Evaluation of a Case-based Facial Action Units Recognition Approach," *2006*

37. M. Khademi and L. -P. Morency, "Relative facial action unit detection,"

38. Z. Liu, R. Liu, Z. Shi, L. Liu, X. Mi and K. Murase, "Semi-Supervised Contrastive Learning with Soft Mask Attention for Facial Action Unit Detection,"

39. M. Nadeeshani, A. Jayaweera and P. Samarasinghe, "Facial Emotion Prediction through Action Units and Deep Learning,"

40. S. Pulido-Castro, N. Palacios-Quecan, M. P. Ballen-Cardenas, S. Cancino-Suárez, A. Rizo-Arévalo and J. M. L. López, "Ensemble of Machine Learning Models for an Improved Facial Emotion Recognition,"

WORK TO BE DONE :

Try to build a better model which performs well on mapping facial action units to emotions.

Try out better mapping of action units to emotions with intensities too.

get a better dataset as DISFA has very few occurrences of action units.

add AU detections into stress module lots of apache 2.0 licenses available

https://paperswithcode.com/task/facial-action-unit-detection