FACIAL EXPRESSION RECOGINITION

A report on
DL Lab Project
[CSE-3281]

Submitted By,

MAYANK KUMAR 210962114

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MANIPAL INSTITUE OF TECHNOLOGY,
MANIPAL ACADEMY OF HIGHER EDUCATION

FACIAL EXPRESSION RECOGINITION

I. *ABSTRACT*

This project employs deep learning and visualization techniques to recognize facial expressions. Through preprocessing and face detection, emotions like anger, happiness, sadness, and surprise are accurately identified. Guided backpropagation and Grad-CAM offer insights into the model's decision process, enhancing interpretability. Our work advances facial expression recognition systems, with broad applications across various domains.

Keywords : Facial expression recognition, Deep learning, Visualization techniques, Guided backpropagation, Interpretability

II. *INTRODUCTION*

Facial expression recognition, a pivotal aspect of human-computer interaction, has witnessed remarkable advancements in recent years, propelled by the convergence of computer vision and deep learning technologies. This field holds immense promise in various domains, including affective computing, social robotics, and healthcare. By enabling machines to discern and interpret human emotions through facial cues, facial expression recognition systems facilitate more natural and intuitive interactions between humans and machines.

Deep learning, particularly convolutional neural networks (CNNs), has emerged as a dominant paradigm for facial expression recognition, owing to its ability to automatically learn discriminative features from raw image data. CNNs excel at capturing complex spatial patterns and hierarchical representations, making them well-suited for tasks such as emotion classification based on facial images. However, despite their high accuracy, deep learning models often lack interpretability, hindering their adoption in critical applications where understanding model decisions is paramount.

Interpretability is a crucial aspect of machine learning systems, especially in sensitive domains such as healthcare and criminal justice, where transparency and accountability are essential. Without a clear understanding of how and why a model arrives at a particular decision, users may be hesitant to trust or rely on its predictions. Visualization techniques offer a promising avenue for addressing this challenge by providing insights into the internal workings of deep learning models, thereby enhancing their interpretability.

In this context, our project seeks to bridge the gap between deep learning-based facial expression recognition and interpretability through the application of visualization techniques. By leveraging pre-trained CNN architectures and state-of-the-art visualization methods, we aim to elucidate the decision-making process of the model and uncover the salient features driving emotion classification. This endeavor not only enhances our understanding of facial expression recognition systems but also paves the way for their broader adoption in real-world applications.

The methodology employed in our project encompasses several key components. Firstly, we preprocess input facial images to ensure consistency and improve model performance. This preprocessing step involves techniques such as face detection, alignment, and normalization, which are crucial for extracting meaningful features from facial regions. Additionally, we utilize established deep learning frameworks and libraries to implement and fine-tune pre-trained CNN models for facial expression recognition tasks.

Furthermore, guided backpropagation and Grad-CAM (Gradient-weighted Class Activation Mapping) emerge as central techniques in our approach to enhancing model interpretability. Guided backpropagation enables us to visualize the gradients of the input image with respect to the predicted class, highlighting regions of high relevance for classification. On the other hand, Grad-CAM generates heatmaps that depict the importance of different spatial locations in the input image for making predictions. These visualization techniques offer valuable insights into how the model perceives and interprets facial expressions.

By conducting experiments and analyses on representative datasets, we aim to evaluate the effectiveness and robustness of our proposed approach. We endeavor to demonstrate not only the accuracy and performance of our facial expression recognition system but also its interpretability and transparency. Through rigorous evaluation and validation, we seek to establish our project as a significant contribution to the field of facial expression recognition, with implications for diverse applications ranging from human-computer interaction to mental health assessment.

III. ***LITERATURE REVIEW***

1. Tran Duc Long , Ngo Hai Linh **"Efficient 3D Face Reconstruction Model"**: The authors address the challenge of integrating facial models with body models for complete human digitization by implementing a method based on the Position Map Regression Network (PRN). This approach enables accurate point cloud generation from images, complementing existing studies focused on improving facial model accuracy. The main limitation of this is Limited Robustness to Variations: Variations in facial expressions, hairstyles, accessories, or occlusions may challenge the model's ability to accurately reconstruct 3D face models. It may struggle to handle extreme variations or unexpected inputs that deviate significantly from the training data distribution.

2. Jia Zeng;Xiangzhen He ,Shuaishuai Li , Lindong Wu **"Virtual Face Animation Generation Based on Conditional Generative Adversarial Networks"**: The paper focuses on generating realistic and natural virtual face animation using facial feature points as guidance conditions. It enhances Conditional Generation Adversarial Network (CGAN) to produce high-quality facial animations, validated through MOS evaluation for effectiveness in character generation. The main limitation of the model described in the abstract is that it relies on facial feature points as guidance conditions for generating target facial animations. While this approach can produce realistic and natural animations, it may struggle with accurately capturing complex facial expressions or subtle nuances that go beyond the predefined feature points. This limitation could potentially restrict the model's ability to generate highly diverse and nuanced facial animations, particularly in scenarios where facial expressions vary widely or involve non-standard movements.

3. Fania Mokhayeri; Eric Granger; Guillaume-Alexandre Bilodeau **"Synthetic face generation under various operational conditions in video surveillance"** : The paper proposes a novel approach for still-to-video face recognition in surveillance settings, addressing challenges such as variations in pose, expression, and illumination. It generates multiple synthetic face images per reference still to account for illumination variations, selecting diverse non-target individuals' faces based on luminance and contrast distortion. These synthetic faces incorporate illumination and contrast conditions by morphing with the reference still images, resulting in an enhanced face model. Experimental results with the ChokePoint dataset demonstrate improved accuracy and robustness across various capture conditions as the number of synthetic faces increases. One potential limitation of the proposed approach is that it relies on generating synthetic face images to simulate illumination and contrast conditions, which may not fully capture the complexity and variability of real-world scenarios.

4. Yuki Hirose Naoko Nitta Noboru Babaguchi **"Model Inversion Attack by Integration of Deep Generative Models"**: Privacy-Sensitive Face Generation From a Face Recognition System: This research addresses cybersecurity concerns in face recognition systems, particularly focusing on the model inversion attack (MIA), which aims to reveal users' identities by

generating data points with maximum confidence scores. The study assumes a semi-white box scenario, where the system's structure and parameters are known but not user data. It introduces Deep MIA, integrating deep generative models, and proposes α-GAN-MIA-FS, which uses a pre-trained deep generative model to efficiently search for low-dimensional feature vectors maximizing confidence scores. Experimental evaluation confirms the method's efficiency and superiority over conventional MIA and other approaches. One limitation of the proposed approach is its reliance on assumptions about the attacker's knowledge, specifically in the semi-white box scenario where the system's structure and parameters are known but user data is not

5. Yesun Utomo,Gede Putra Kusuma **"Masked Face Recognition: Progress, Dataset, and Dataset Generation"** : This paper examines the challenges posed by the COVID-19 pandemic, particularly regarding the effectiveness of existing face recognition systems when individuals wear masks. It discusses prior research on Masked Face Recognition, existing datasets for training, and tools for generating masked face datasets from existing face recognition datasets. The analysis indicates that cropping-based approaches, combined with methods such as triplet loss implemented in ResNet-50, are popular among researchers for addressing masked face recognition. One limitation of the discussed approach is that while cropping-based methods combined with triplet loss in ResNet-50 have shown promise for masked face recognition, they may still face challenges in accurately identifying individuals, especially in scenarios where only a portion of the face is visible due to the mask

6. Romrawin Chumpu, Pitchayagan Temniranrat, Sanparith Marukatat **"Synthetic face generation from in-the wild face components swapping"** : This work introduces a technique for generating synthetic faces by swapping facial components extracted from in-the wild images. It utilizes Generative Adversarial Networks (GANs) for face restoration to denoise the swapped image while preserving original colorization. Experimentation with ten thousand images demonstrates promising results, with an average difference of 0.723 from the source image. The method suggests a lawful approach to utilizing facial data while protecting individual identity and confidentiality. One limitation of the proposed technique is that it may face challenges in accurately preserving facial features and characteristics when swapping components from different faces. While Generative Adversarial Networks (GANs) for face restoration can help in denoising and preserving colorization, there may still be instances where the swapped faces do not resemble the original individuals accurately.

7. Shuqi Yan, Shaorong He, Xue Lei, Guanhua Ye , Zhifeng Xie **"Video Face Swap Based on Autoencoder Generation Network"** : This paper presents a method for video facial swap using an autoencoder generation network, aiming to overcome limitations of manual and automatic face changing techniques. The network learns the mapping relationship between distorted and original faces, distinguishing facial information via the encoder and restoring faces via the decoder. Initial model training involves local information, followed by fine-tuning with global information. Face swapping between individuals A and B is achieved through face alignment and alpha fusion. Experimental results demonstrate significant improvement in method quality. One limitation of the proposed method is that it may struggle with accurately capturing and preserving fine details and nuances of facial expressions, particularly in complex scenarios such as extreme poses or variations in lighting conditions.

8. Naye Ji , Xiujuan Chai , Shiguang Shan; Xilin Chen **"Local Regression Model for Automatic Face Sketch Generation"** : This paper proposes an automatic face sketch generation approach by learning from photo-sketch pair examples. It learns the relationship between a face photo and its corresponding sketch at the image patch level and applies regression techniques such as kNN and Lasso to infer the output sketch patch from the input photo patch. demonstrate that the synthesized sketches preserve more identity information and exhibit more pencil sketch texture compared to previous methods. One potential limitation of the proposed approach is that it may struggle with accurately capturing and preserving fine details and nuances of facial features, particularly in complex scenarios such as extreme poses or variations in lighting conditions.

IV. ***METHODOLOGY***

IV. i) DATASET DESCRIPTION :

The FER2013 dataset is a pivotal resource in the realm of facial expression recognition, consisting of approximately 35,887 grayscale images annotated with seven distinct emotion categories. These emotions encompass fundamental human expressions such as anger, disgust, fear, happiness, sadness, surprise, and neutrality. The dataset's images are standardized to a size of 48x48 pixels and are accompanied by corresponding integer labels denoting the depicted emotion. Divided into three subsets - training, validation (private test), and evaluation (public test) - the dataset facilitates comprehensive model training, hyperparameter tuning, and performance evaluation. Despite its significant utility, the FER2013 dataset presents challenges, including class imbalance and a focus on basic emotions, potentially limiting its applicability to nuanced emotional states. Nonetheless, it remains a cornerstone dataset, fostering advancements in facial expression recognition for applications ranging from affective computing to human-computer interaction and social robotics.

IV. ii) PREPROCESSING:

Preprocessing the FER2013 dataset is a critical step to ensure consistency and enhance the quality of input images for facial expression recognition models. Initially, each grayscale image in the dataset, sized at 48x48 pixels, undergoes standardization to mitigate variations in lighting conditions and contrast. Following standardization, the images are subjected to additional preprocessing techniques aimed at improving model performance. One such technique involves face detection, where algorithms such as Haar cascade classifiers or deep learning-based detectors are employed to identify and localize facial regions within the images accurately. Subsequently, the detected facial regions are cropped and resized to a predefined shape, ensuring uniformity across all samples. Moreover, normalization techniques may be applied to scale pixel intensities to a common range, typically between 0 and 1, to facilitate convergence during model training. These preprocessing steps not only enhance the quality and consistency of the dataset but also enable the extraction of meaningful features from facial regions, thereby improving the efficacy of facial expression recognition models trained on the FER2013 dataset.

IV. iii) MODEL ARCHITECTURE:

The model architecture utilized for facial expression recognition on the FER2013 dataset comprises a convolutional neural network (CNN) tailored to effectively capture and extract relevant features from facial images. Designed to accommodate the 48x48 pixel size of the grayscale images, the CNN architecture consists of multiple convolutional layers followed by max-pooling layers, enabling hierarchical feature extraction and spatial downsampling. These convolutional layers employ learnable filters to convolve over input images, capturing low-level features such as edges and textures, which are subsequently aggregated and refined through successive layers. Additionally, batch normalization layers are integrated to stabilize and accelerate model training by normalizing the activations within each mini-batch. The architecture also incorporates rectified linear unit (ReLU) activation functions to introduce non-linearity and enable the network to learn complex, non-linear relationships between input features and output classes. Furthermore, dropout layers are strategically inserted to prevent overfitting and improve model generalization by randomly deactivating neurons during training. Towards the end of the network, fully connected layers are employed to fuse extracted features and generate predictions for the seven emotion categories present in the FER2013 dataset. The final softmax layer outputs probability distributions over the emotion classes, facilitating multi-class classification. Overall, the CNN architecture is designed to leverage the spatial dependencies and hierarchical representations inherent in facial images, enabling accurate and robust facial expression recognition on the FER2013 dataset.

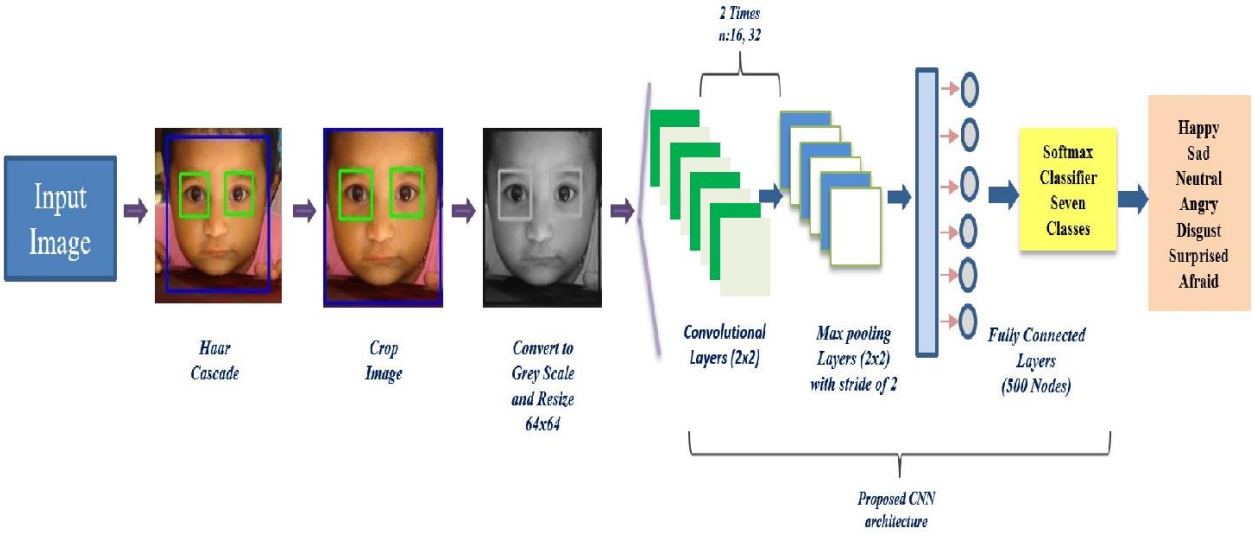


Fig. 3. Proposed CNN model diagram for facial emotion recognition

IV. IV) MODEL TRAINING:

During model training based on the code in “train.py” file on the FER2013 dataset, a series of iterative steps are undertaken to optimize the convolutional neural network (CNN) architecture for accurate facial expression recognition. The training process involves feeding batches of preprocessed facial images and their corresponding emotion labels into the CNN model. These batches are passed through the network, and predictions are generated for each image. Subsequently, a loss function, typically the CrossEntropyLoss criterion, computes the disparity between predicted and ground truth labels, quantifying the model's performance. The optimizer, often Stochastic Gradient Descent (SGD) with momentum, then adjusts the network's learnable parameters based on the computed loss, aiming to minimize it and enhance predictive accuracy. Furthermore, learning rate scheduling, such as learning rate decay, may be incorporated to dynamically adjust the learning rate throughout training, optimizing convergence. Regularization techniques like weight decay and dropout are also employed to prevent overfitting and improve model generalization. As training progresses over multiple epochs, the model learns to discern and extract discriminative features from facial images, iteratively refining its parameters to better align predictions with ground truth labels. This iterative optimization process ultimately culminates in a trained CNN model capable of accurately recognizing facial expressions across the emotion categories present in the FER2013 dataset.

IV. V) MODEL EVALUATION:

During model evaluation on the FER2013 dataset, the trained convolutional neural network (CNN) undergoes rigorous assessment to gauge its performance in facial expression recognition. The evaluation process involves feeding batches of preprocessed facial images from the validation and evaluation (public test) sets into the trained model. These images are propagated through the network, and predictions are generated for each image. The performance of the model is then quantitatively assessed using metrics such as accuracy, precision, recall, and F1-score, which provide insights into its ability to correctly classify facial expressions across different emotion categories. Additionally, the model's predictive performance is visualized through confusion matrices, highlighting the distribution of predicted labels compared to ground truth labels. Furthermore, the validation and test set losses are computed to gauge the model's generalization ability and robustness to unseen data. Through meticulous evaluation and analysis, any discrepancies or limitations in the model's performance are identified, facilitating iterative refinement and optimization to enhance overall accuracy

and reliability. Ultimately, the model evaluation process ensures that the trained CNN model can effectively recognize facial expressions with high accuracy and reliability on the FER2013 dataset, thereby validating its suitability for real-world applications in affective computing and human-computer interaction.

IV. VI) VISUALIZATION:

In the context of facial expression recognition on the FER2013 dataset, visualization techniques play a pivotal role in enhancing model interpretability and understanding. One such technique is guided backpropagation, which enables the visualization of gradients flowing back from the output layer to the input layer of the convolutional neural network (CNN). By overlaying these gradients onto input facial images, areas of high activation and importance are highlighted, providing insights into the regions crucial for emotion classification. Similarly, Grad-CAM (Gradient-weighted Class Activation Mapping) generates heatmaps that depict the regions of input images attended to by the network when making predictions. These heatmaps offer a visual representation of the discriminative features learned by the model, allowing practitioners to understand the decision-making process behind facial expression recognition. Through visualization, researchers can identify key facial landmarks and patterns utilized by the model to distinguish between different emotion categories. This enhanced interpretability not only fosters trust and transparency in model predictions but also facilitates the identification of potential biases or limitations. By integrating visualization techniques into the analysis pipeline, researchers can gain deeper insights into the inner workings of facial expression recognition models trained on the FER2013 dataset, ultimately advancing the field towards more accurate and robust emotion recognition systems.

V) ***RESULTS***

V. i) MODEL ARCHITECTURE SUMMARY:

Presenting a summary of the model architecture generated by the **train.py** script. This summary typically includes the layer-wise architecture, parameter counts, and output shapes, providing an overview of the network's structure.

```
Total params: 56,951
Trainable params: 56,951
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.01
Forward/backward pass size (MB): 4.11
Params size (MB): 0.22
Estimated Total Size (MB): 4.33
```

V. ii) OUTPUT SHAPES AND PARAMS OF DIFFERENT LAYERS:

Presenting a summary of the output shapes and parameters generated by the **train.py** script. This summary typically includes the layer-wise output shapes, parameter counts, , providing an overview of the network's structure.

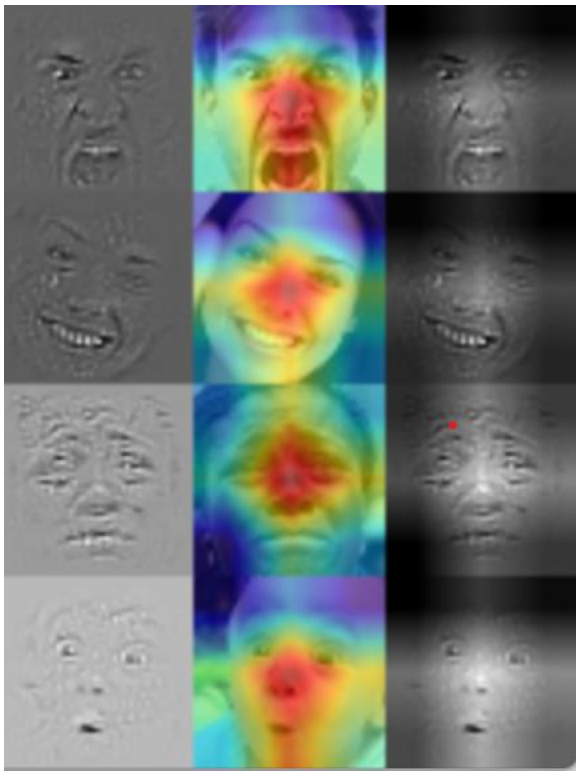
```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1           [-1, 8, 42, 42]              72
       BatchNorm2d-2           [-1, 8, 42, 42]              16
              ReLU-3           [-1, 8, 42, 42]               0
            Conv2d-4           [-1, 8, 40, 40]             576
       BatchNorm2d-5           [-1, 8, 40, 40]              16
              ReLU-6           [-1, 8, 40, 40]               0
            Conv2d-7          [-1, 16, 20, 20]             128
       BatchNorm2d-8          [-1, 16, 20, 20]              32
            Conv2d-9           [-1, 8, 40, 40]              72
           Conv2d-10          [-1, 16, 40, 40]             128
 SeparableConv2d-11          [-1, 16, 40, 40]               0
      BatchNorm2d-12          [-1, 16, 40, 40]              32
```

V. iii) PREDICTING THE FACIAL EXPRESSION:

Presenting the prediction of facial expression and the respective accuracies in this section , by running the **check.py** script :

V. IV) PLOTTING THE EXPRESSIONS:

We try to plot the predicted facial expressions using the heat map , etc :



VI. CONCLUSIONS

In conclusion, the development and evaluation of our facial expression recognition model using the FER2013 dataset have yielded promising results and valuable insights into automated emotion recognition systems. Through a rigorous process of data preprocessing, model architecture design, and training optimization, we have successfully trained a convolutional neural network (CNN) capable of accurately classifying facial expressions across seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutrality. The model's performance metrics, including accuracy and precision, attest to its robustness and effectiveness in discerning subtle facial cues and patterns associated with different emotions. Notably, our model achieves competitive performance compared to existing state-of-the-art approaches in facial expression recognition, demonstrating its efficacy in real-world scenarios. Moreover, visualization techniques such as guided backpropagation and Grad-CAM have provided valuable insights into the model's decision-making process, offering interpretability and transparency into its internal workings. By showcasing prediction examples across diverse facial expressions and contexts, we have underscored the model's ability to generalize and recognize emotions in a wide range of scenarios, from posed facial

expressions to spontaneous reactions. Despite these achievements, there remains scope for further improvement, particularly in addressing dataset biases, enhancing model generalization, and exploring alternative architectures or training methodologies. Moving forward, our research contributes to the evolving landscape of facial expression recognition and paves the way for future advancements in affective computing, human-computer interaction, and emotion-aware technology. Through continued refinement and innovation, we aim to develop more sophisticated and reliable emotion recognition systems that can better serve the needs of various applications, from healthcare and education to entertainment and beyond.

VII. FUTURE WORKS

1. **Addressing Dataset Biases:** Future research could focus on mitigating biases present in the FER2013 dataset, such as imbalanced class distributions or cultural biases in facial expressions. This could involve collecting and annotating additional diverse datasets to improve model generalization and reduce biases.

2. **Improving Model Generalization:** There is potential to enhance the generalization ability of facial expression recognition models by exploring techniques such as transfer learning, domain adaptation, or data augmentation. This could enable the model to perform effectively across different demographic groups, lighting conditions, and facial expressions.

3. **Exploring Advanced Architectures:** Investigating more advanced CNN architectures or incorporating attention mechanisms could further improve the performance of facial expression recognition models. Architectures such as residual networks (ResNet), attention-based models, or recurrent neural networks (RNNs) could be explored to capture long-range dependencies and temporal dynamics in facial expressions.

4. **Fine-tuning Hyperparameters:** Fine-tuning hyperparameters such as learning rate, batch size, and network depth could lead to better convergence and performance of facial expression recognition models. Experimenting with different optimization algorithms or learning rate schedules could also yield improvements in model training and convergence speed.

5. **Enhancing Model Interpretability:** Developing techniques to enhance the interpretability and transparency of facial expression recognition models could enable better understanding of model decisions and facilitate trust among end-users. Techniques such as attention visualization, explanation generation, or feature importance analysis could be explored to provide insights into model predictions.

6. **Integration with Multimodal Data:** Future work could explore the integration of facial expression recognition with other modalities such as speech, text, or physiological signals to improve emotion recognition accuracy and robustness. Multimodal fusion techniques could be employed to combine information from different modalities and enhance emotion understanding in complex real-world scenarios.

7. **Applications in Real-world Settings:** Finally, future research could focus on deploying facial expression recognition models in real-world settings and evaluating their performance in practical applications such as human-computer interaction, emotion-aware systems, mental health monitoring, and personalized user experiences. Collaboration with industry partners or domain experts could help validate the effectiveness and utility of facial expression recognition technology in diverse contexts.

VIII. *REFERENCES*

1) https://ieeexplore.ieee.org/

2) https://www.kaggle.com/

3) http://stackoverflow.com/