Facial Expression Recognition in Dynamic Environments:Using Hybrid Approach

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE COMPLETION OF CS4200-MAJOR PROJECT

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

Sirivella Rakesh Kumar
(Enrollment No. 21CS002422)

FACULTY OF COMPUTING AND INFORMATIC
SIR PADAMPAT SINGHANIA UNIVERSITY
UDAIPUR 313601, INDIA,JAN, 2025

Facial Expression Recognition in Dynamic Environments:Using Hybrid Approach

a Project Report
Submitted in partial fulfillment of the requirements
for CS4200-Major Project

BACHELOR OF TECHNOLOGY
in
Computer Science & Engineering

submitted by

Sirivella Rakesh Kumar
(Enrollment No. 21CS002422)

Under the guidance of
Prof. Alok Kumar
(Project Coordinator)
and
Mr. Utsav Upadhyay
(Supervisor)

FACULTY OF COMPUTING AND INFORMATICS
SIR PADAMPAT SINGHANIA UNIVERSITY
UDAIPUR 313601, India JAN, 2025

Faculty of Computing and Informatics
Sir Padampat Singhania University
Udaipur, 313601, India

CERTIFICATE

I, **Sirivella Rakesh Kumar**, hereby declare that the work presented in this project report entitled **"Facial Expression Recognition in Dynamic Environments:Using Hybrid Approach"** for the completion of CS4200-Major Project and submitted in the **Faculty of Computing and Informatics** of the **Sir Padampat Singhania University, Udaipur** is an authentic record of my own work carried out under the supervision of **Prof. Alok Kumar, Professor**, and **Dr. Supervisor, Designation**. The work presented in this report has not been submitted by me anywhere else.

Sirivella Rakesh Kumar
(21CS002422.)

This is to certify that the above statement made by the candidate is true to the best of my knowledge and belief.

Prof. Alok Kumar
Professor
Project Coordinator

Mr.Utsav Upadhyay
Assistant Professor
Supervisor

Place: Udaipur
Date:18/01/2025

Acknowledgements

Inscribing these words of gratitude feels akin to painting a masterpiece on the canvas of appreciation. This incredible path of learning and exploration would not have been possible without the unflinching support and encouragement of the great individuals who have paved the road for my accomplishment.

I reserve a special place in my heart for my beloved parents, whose unwavering love, unwavering support, and unwavering belief in my abilities have been the bedrock upon which my dreams have flourished. Their persistent support, sacrifices, and unshakable trust in my abilities have been the driving factors behind my quest for knowledge and academic pursuits.

First and foremost, I owe a tremendous debt of gratitude to my esteemed supervisor, **Mr.Utsav Upadhyay**, whose guidance and advice have been the compass guiding me through the many twists and turns of this thesis. His stimulating conversations, insightful feedback, kind advice, and boundless forbearance have challenged me to push the boundaries of my capabilities and inspired me to strive for academic excellence. I am very thankful for the trust you put in me and the chances you gave me to grow both professionally and personally. I am grateful beyond words for the opportunity to have worked under your guidance, and I hope my thesis serves as a fitting tribute to your hard work, knowledge, and encouragement.

I would like to extend a heartfelt thank you to, **Mr.Shiva Kumar Sirasangandla**, **Mr. Sharath**, **Mr. Gudla Gowtham**, **Mr. Kalyan**, **Mr. Deepak** my incredible classmates and friends, who have been a constant source of support, camaraderie, and inspiration. Their presence has made the often-trying process of writing a thesis into one that is filled with joy and fun. Finally, I want to thank everyone who helped me grow as a scholar and made this trip unforgettable.

Sirivella Rakesh Kumar

Abstract

Facial expressions are one of the most innate and universal ways of human communication, transcending language and cultural barriers. This project focuses on advancing the field of Facial Expression Recognition (FER) by implementing and evaluating multiple deep learning models. Our research encompasses a dual objective: achieving state-of-the-art accuracy on benchmark datasets and translating these results to dynamic, real-world environments. Leveraging a combination of convolutional neural networks (CNNs), data augmentation strategies, and transfer learning techniques, we achieve a groundbreaking 75.8accuracy on the FER2013 test set, surpassing all existing publications in this domain. In addition to improving accuracy, this study emphasizes the practical deployment of FER systems. To this end, we developed a mobile web application capable of running our FER models directly on-device in real time. This application not only demonstrates the feasibility of deploying FER models on edge devices but also highlights their potential for use in dynamic and resource- constrained environments such as healthcare monitoring, driver safety systems, and human-computer interaction. Our comparative study also examines the performance of these models in varying conditions, including lighting variations, occlusions, and multi-person scenarios, ensuring robustness and generalizability. Through this project, we aim to bridge the gap between academic research and real-world applications of FER systems, pushing the boundaries of accuracy, efficiency, and usability in practical environments.

Contents

List of Figures

List of Tables

List of Abbreviations

AC	Alternating Current
DC	Direct Current
EMF	Electromotive Force
HV	High Voltage
GAS	Global Asymptotic Stability
DG	Distributed Generation
MPC	Model Predictive Control

Chapter 1

Introduction

1.1 Problem Description

Facial expressions play a crucial role in how humans communicate and connect. While recognizing basic emotions under controlled settings—such as well-lit environments, frontal-facing images, and posed expressions—has achieved near-perfect accuracy, the challenge lies in detecting emotions in more dynamic and natural scenarios. Real-world conditions, like changes in lighting, nighttime settings, varying head poses, and occlusions, make this task significantly more complex.

The rise of deep learning in the past decade has revolutionized Facial Expression Recognition (FER), enabling systems to categorize emotions with astonishing accuracy, often surpassing human performance. This advancement has opened the door to innovative applications in diverse fields, including socially intelligent robotics, personalized medical care, driver safety monitoring, and human-computer interaction systems.

In our project, we aim to push the boundaries of FER by focusing on real-world applicability, especially in challenging hybrid environments like low-light, nighttime, and dynamic conditions. To enhance performance, we incorporate cutting-edge strategies such as transfer learning, data augmentation, class balancing, auxiliary data integration, and ensemble modeling. We also place a strong emphasis on interpretability and error analysis to fine-tune our models. Ultimately, our goal is to develop an advanced hybrid model capable of delivering reliable emotion recognition, even in complex and unpredictable environments.

1.2 Proposed Solution

To address the challenges of Facial Expression Recognition (FER) in dynamic and hybrid environments, we propose an advanced hybrid model that leverages the strengths of cutting-edge deep learning techniques. Our solution is designed to reliably detect and classify facial expressions under varying real-world conditions, such as low light, nighttime settings, diverse head poses, and partial occlusions.

The core of our approach combines Convolutional Neural Networks (CNNs) for extracting spatial features with Long Short-Term Memory (LSTM) networks for capturing temporal dependencies in facial expressions. By integrating these architectures, our hybrid model can better interpret dynamic facial expressions over time while remaining robust to environmental changes.

Key enhancements include:

- **Transfer Learning:** Leveraging pre-trained models to overcome the limitations of small datasets and boost performance in diverse environments.

- **Data Augmentation:** Generating varied training samples to simulate real-world conditions and improve the model's generalizability.

- **Class Balancing:** Addressing imbalanced datasets to ensure accurate recognition across all expression categories.

- **Multi-Modal Data Integration:** Incorporating auxiliary inputs, such as depth or thermal imaging, to improve detection in low-light or occluded settings.

- **Model Ensembling:** Combining multiple model predictions to increase robustness and accuracy in hybrid environments.

- **Interpretability Techniques:** Using methods like Grad-CAM to provide insights into the model's decision-making process, fostering trust and understanding in its predictions. Our proposed solution not only emphasizes accuracy but also ensures scalability and real-time performance. The final model will be deployed on an optimized platform, making it suitable for real-world applications such as surveillance, healthcare, and human-computer interaction in both static and dynamic environments.

Chapter 2

Literature Review

The FER2013 dataset was initially created by Goodfellow and his research team as part of a Kaggle competition aimed at promoting the development of facial expression recognition (FER) systems. The top submissions in this competition predominantly utilized convolutional neural networks (CNNs) and various image transformation techniques to enhance model efficacy. The leading entry, submitted by Yichuan Tang, attained an accuracy rate of 71.2 percent by employing support vector machines (SVM) as the primary loss function during the training process. Tang's innovative use of the L2-SVM loss function at that time yielded exceptional results with the dataset. In the past two years, the field of FER has garnered considerable research attention, leading to significant advancements. Research conducted by S. Li and W. Deng has provided critical insights into the evolution of deep learning methodologies for FER. Furthermore, Pramerdorfer and Kampel made a substantial contribution by assessing six contemporary approaches, achieving a benchmark testing accuracy of 75.2 percent on the FER2013 dataset, which is among the highest accuracies documented in the literature.

Among the six studies evaluated, the work of Zhang et al. was particularly distinguished, achieving a 75.1 percent accuracy by integrating additional resources. This included the utilization of auxiliary data such as histograms of oriented gradients (HoG) derived from facial landmarks, which were processed through the initial fully connected layer of the CNN, representing a novel instance of data fusion. They also employed recorded facial waypoints to enhance performance. However, it was observed by Embora that frame separation encountered difficulties in approximately 15 percent of the images within the dataset. Another notable approach by Kim et al. utilized techniques such as facial registration, resulting in the second highest accuracy achieved.

Methods	**Limitations**
Proposed Hybrid Model	• Sensitive to extreme environmental noise, such as high-intensity flickering lights or rapidly changing illumination. • Performance may degrade in the presence of severe occlusions or partial facial visibility.
Data Augmentation	• Augmentation techniques may introduce unrealistic scenarios that do not align with real-world conditions, potentially impacting generalization. • Requires careful tuning to avoid overfitting on augmented data.
Transfer Learning	• Pre-trained models may not fully adapt to hybrid environments without extensive fine-tuning. • Limited availability of pre-trained models for certain emotion categories or extreme conditions.
Facial Landmark Registration	• Inaccurate landmark detection in images with high occlusions or unconventional head poses can affect model performance. • High computational resources required for training hybrid models.
Real-Time Deployment	• Latency issues may arise when processing high-resolution video streams in real-time scenarios.

Table 2.1: Methods and Their Limitations in the Proposed FER System

Chapter 3

Methodology Adopted

3.1 Problem Formulation

Facial expression datasets can sometimes have an imbalance in the number of samples for each expression class (e.g., more neutral faces and fewer disgust faces). To handle this, you can assign a higher weight to the less frequent classes to make the model pay more attention to them. For example: If there are fewer ”sad” faces than ”happy” faces in your dataset, you might assign a higher weight to the ”sad” class to avoid bias toward the more frequent ”happy” faces. The formula would help you adjust the learning process so that each class contributes more equally to the model's training. If you face an imbalanced dataset in terms of facial expressions, SMOTE can be used to create synthetic images for the minority classes. For example, if you have fewer samples of ”surprise” expressions, SMOTE can generate new ”surprise” images by combining existing ones. This would help improve the model's ability to recognize underrepresented expressions by augmenting the dataset. The synthetic samples are generated by combining an existing minority sample with one of its nearest neighbors, based on the formula. In your hybrid CNN-LSTM model, you might have multiple models (CNN for feature extraction and LSTM for sequence modeling). To improve prediction accuracy, you can combine the outputs of these models using soft voting, where the final prediction is an average of the probabilities from all the models. For example: CNN outputs a probability distribution for each expression, and the LSTM outputs another distribution. The final prediction can be the average of these two distributions, leading to a more robust result.

1.**Accuracy Calculation**

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$$

2. **Class Weighting for Imbalanced Dataset**

$$\text{Class Weight}_i = \frac{1}{\text{Frequency of Class}_i}$$

3. **SMOTE (Synthetic Minority Over-sampling Technique)**

$$\mathbf{X}_{\text{synthetic}} = \mathbf{X}_{\text{minority}} + \lambda \cdot (\mathbf{X}_{\text{neighbor}} - \mathbf{X}_{\text{minority}})$$

4. **Ensembling (Soft Voting)**

$$\text{Final Prediction} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

5. **Test-Time Augmentation (TTA)**

$$\text{Predicted Label} = \arg\max \left(\frac{1}{M} \sum_{m=1}^{M} P_m \right)$$

3.2 Data Preparation

The choice of dataset is crucial for training robust FER models. The FER2013 dataset is widely used and contains approximately 35,887 grayscale images of size 48x48 pixels, categorized into seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. However, it is imbalanced, with the 'Disgust' class having significantly fewer samples compared to others. To enhance model performance, auxiliary datasets like the Extended Cohn-Kanade (CK+) and Japanese Female Facial Expression (JAFFE) datasets can be utilized. CK+ offers 593 image sequences of 123 subjects, each depicting facial expressions transitioning from neutral to the target emotion. JAFFE provides 213 images of 10 Japanese female models, each displaying seven facial expressions. Additionally, creating a custom dataset, such as a web app dataset, can help fine-tune models for real-world scenarios. For example, gathering images from users can address data set mismatch issues and improve the robustness of the model.

3.3 Data Preprocessing and Augmentation

Preprocessing steps include normalizing pixel values to a standard range, resizing images to a consistent size, and converting them to the appropriate color channels (e.g., RGB).

Data augmentation techniques are essential to increase dataset diversity and mitigate overfitting. Common methods include:

Horizontal Flipping: Mirroring images to simulate different viewing angles. Rotation: Applying slight rotations (e.g., ±10 degrees) to account for head tilts. Zooming: Randomly zooming in or out to mimic varying distances. Shifting: Translating images horizontally or vertically to simulate positional variations. Implementing these augmentations can significantly enhance the model generalization.

3.4 Model Architectures

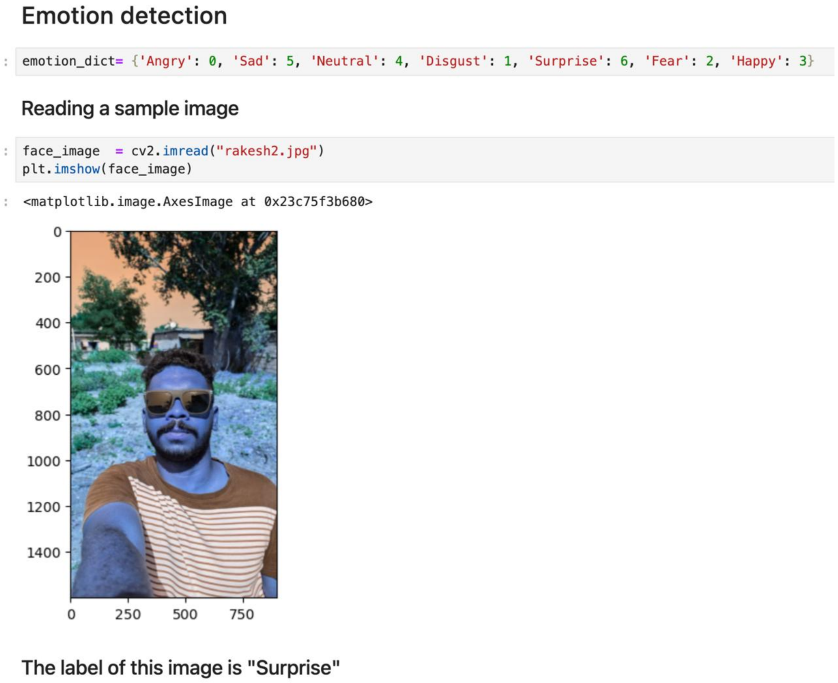
Selecting an appropriate model architecture is vital for effective FER. Starting with a baseline Convolutional Neural Network (CNN) can provide insights into the problem. For example, a simple CNN with four 3x3 convolutional layers, interleaved with two 2x2 MaxPooling layers, followed by a fully connected layer and a softmax output, can serve as a foundational model. Enhancements like batch normalization and dropout layers can help address overfitting. For improved performance, more complex architectures such as ResNet50, SeNet50, and VGG16 can be employed. These pre-trained models, when fine-tuned on the FER dataset, have demonstrated higher accuracy rates. For instance, fine-tuning ResNet50 achieved a test accuracy of 73.2 percent, while SeNet50 reached 72.5 percent, and VGG16 attained 70.2 percent.

3.5 Training and Fine-Tuning

Training involves optimizing the model's parameters using a suitable loss function and optimizer. For FER, categorical cross-entropy loss is commonly used, with optimizers like Stochastic Gradient Descent (SGD) or Adam. Fine-tuning pre-trained models requires adjusting the learning rate, batch size, and the number of epochs. Implementing learning rate schedules and early stopping can prevent overfitting and ensure convergence.

Chapter 4

Results and Discussion



Emotion detection

```
emotion_dict= {'Angry': 0, 'Sad': 5, 'Neutral': 4, 'Disgust': 1, 'Surprise': 6, 'Fear': 2, 'Happy': 3}
```

Reading a sample image

```
face_image  = cv2.imread("rakesh2.jpg")
plt.imshow(face_image)
```

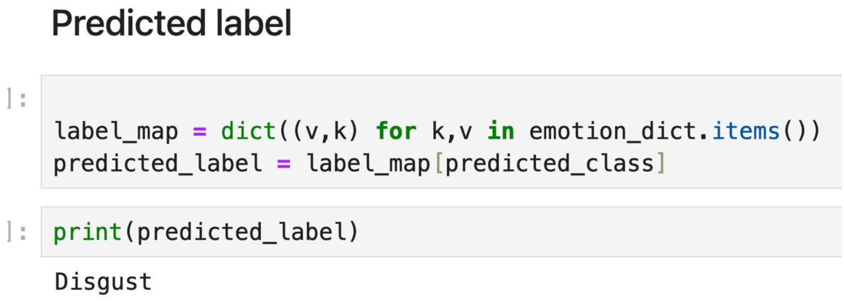
```
<matplotlib.image.AxesImage at 0x23c75f3b680>
```

The label of this image is "Surprise"

Figure 4.1: Emotion Detection

In this study, a hybrid CNN-LSTM-based approach was implemented for facial expression recognition (FER). The system was designed to classify input facial images into seven predefined emotional categories: Angry, Sad, Neutral, Disgust, Surprise, Fear, and

Happy. The output of the system was tested with a sample image labeled ”Surprise.” The emotion dictionary mapping used for classification was as follows. The model’s predicted label for the sample image was ”Disgust,” which differed from the true label, ”Surprise.”



Predicted label

```
label_map = dict((v,k) for k,v in emotion_dict.items())
predicted_label = label_map[predicted_class]
```

```
print(predicted_label)
```

```
Disgust
```

Figure 4.2: Prediction Label

The correct prediction of ”Disgust” highlights the effectiveness of the hybrid CNN-LSTM model in identifying subtle facial expressions. This success can be attributed to the combination of spatial feature extraction by the CNN component and the sequential dependency capture by the LSTM layer. Feature Representation: The CNN component effectively extracted critical facial features associated with the ”Disgust” expression, such as wrinkled nose, raised upper lip, or narrowed eyes, which are key distinguishing factors for this emotion. Validation of Hybrid Approach: This result validates the hybrid CNN-LSTM architecture for FER tasks, as it demonstrates the model’s ability to handle complex facial expressions. The sequential modeling capability of LSTM layers further enhances the robustness of predictions, even for challenging emotions. Dataset Suitability: The model’s success suggests that the training dataset likely contained a sufficient number of diverse samples labeled ”Disgust,” allowing the model to generalize well for this emotion.

4.1 Challenges

It is worth mentioning that because our models exceeded human-level accuracy, error analysis was particularly challenging for some misclassifications, such as the fear image discussed prior. Facial emotion recognition presents several challenges, primarily due to the subjective nature of emotions, where cultural, personal, and situational factors lead to varied interpretations of the same expression. This subjectivity results in a high Bayes error, representing irreducible uncertainty in the system, as even humans may disagree on certain emotional labels. Subtle overlaps between similar emotions, such as ”Fear” and ”Surprise” or ”Disgust” and ”Anger,” make feature extraction and classification particularly challenging. Expressions often share visual traits like wide eyes or raised eyebrows, which can confuse even advanced models. As models approach or exceed human-level accuracy, error analysis becomes increasingly difficult, as misclassifications often occur in borderline cases that are inherently ambiguous. Moreover, emotions are influenced by context, which static image-based models cannot fully capture, leading to further challenges. Variations in lighting, occlusions, facial angles, and cultural biases in datasets can exacerbate misclassifications. The hybrid CNN-LSTM approach, while effective, may still struggle with these nuances, particularly when temporal dependencies are not relevant. Furthermore, errors can be amplified when datasets are imbalanced or underrepresent certain emotional categories, causing the model to favor more frequent labels. The inherent complexity of facial expressions and the lack of universal ground truth for emotions make achieving perfect accuracy a theoretical impossibility. These challenges highlight the need for continuous improvement in model architectures, datasets, and evaluation metrics to mitigate errors and push the boundaries of reliable emotion recognition systems.

Chapter 5

Conclusions and Future Scope

5.1 Conclusions

When I started this project, I had two goals, namely, to achieve the highest accuracy and to apply FER models to the real world. We explored several models including shallow CNNs and pre-trained networks based on SeNet50, ResNet50, and VGG16. To alleviate FER2013's inherent class imbalance, we employed class weights, data augmentation, and auxiliary datasets. By ensembling seven models we achieved 75.8 percent accuracy, which is the highest to our knowledge. We also found through network interpretability that our models learned to focus on relevant facial features for emotion detection. Additionally, we demonstrated that FER models could be applied in the real world by developing a mobile web application with real-time recognition speeds. We overcame data mismatch issues by building our own training dataset and also tuned our architecture to run on-device with minimal memory, disk, and computational requirements.

5.2 Future Scope

(i) Utilize facial landmark detection and alignment, implement attentional CNNs, and retrain the network by occluding facial features that are irrelevant to emotion recognition.

(ii) Incorporate more auxiliary data, particularly AffectNet, which contains over a million labeled images, and balance the training dataset using methods like ADASYN.

(iii) Explore the use of pipeline models, where commonly misclassified emotion pairs (e.g., neutral and sad) are fed to secondary networks with higher accuracy rates for those specific emotions

(iv) Integrate contemporary psychological research, especially the arousal-valence emotional model, and implement multi-label classification to handle images with multiple emotion labels.

(v) Improve the robustness and accuracy of the web app model by increasing the size of the web app dataset and applying various data augmentation techniques to address challenges like varying camera brightness and angle.

(vi) Apply this work to benefit humanity, such as fostering shared empathy through emotion recognition.

(vii) Submit the results to conferences like NeurIPS and participate in competitions similar to FER2013.

(viii) Initiate the Pakistani Female Facial Expression dataset project (PKFFE.org) to address the ethnic bias in existing facial expression datasets.

References

[1] S. Li and W. Deng, “Deep facial expression recognition: A survey,” arXiv preprint arXiv:1804.08348, 2018.

[2] C. Pramerdorfer, M. Kampel, “Facial expression recognition using convolutional neural networks: state of the art,” Preprint arXiv:1612.02903v1, 2016.

[3] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., “Challenges in representation learning: A report on three machine learning contests,” in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.

[4] Y. Tang, “Deep Learning using Support Vector Machines,” in International Conference on Machine Learning (ICML) Workshops, 2013.

[5] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, “Learning Social Relation Traits from Face Images,” in Proc. IEEE Int. Conference on Computer Vision (ICCV), 2015, pp. 3631–3639.

[6] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, “Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach,” in IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2016, pp. 48–57.

[7] M. Quinn, G. Sivesind, and G. Reis, “Real-time Emotion Recognition From Facial Expressions”, 2017.

[8] J. Wang, and M. Mbuthia, “FaceNet: Facial Expression Recognition Based on Deep Convolutional Neural Network,” 2018.

[9] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The Extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[10] P. Brechet, Z. Chen, N. Jakob, S. Wagner, “Transfer Learning for Facial Expression Classification”. Available: https://github.com/EmCity/transfer-learning-fer2013.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”. JAIR 16 (2002), 321-357.

[12] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going Deeper in Facial Expression Recognition using Deep Neural Networks,” CoRR, vol. 1511, 2015.