FINAL REPORT

Akshaj Rakhecha
Student# 1008942174
akshaj.rakhecha@mail.utoronto.ca

Akshay Ravikumar
Student# 1008716294
akshay.ravikumar@mail.utoronto.ca

Aradhya Dang
Student# 1008745379
aradhya.dang@mail.utoronto.ca

Gurmann Singh Jaggi
Student# 1008735488
gurmann.jaggi@mail.utoronto.ca

ABSTRACT

Our final report starts off by giving a brief introduction of the project. This is followed by an illustration along with the model architecture, and its quantitative and qualitative results. We further discuss the implications of the results and evaluate it on new data. Finally, we talk about the project difficulty and potential improvements. —-Total Pages: 9

1 INTRODUCTION

The primary goal of this project is to develop an emotion recognition model that analyzes facial expressions for categorization in real-time. The integrated system comprises a YOLO model for face detection and a custom-designed CNN for highly accurate emotion classification.

The project is motivated by the need to provide accurate interpretations of different facial expressions, predominantly in digital or non-verbal contexts, ranging from security and surveillance systems to detect potential threats to marketing in recognizing customer responses to products and advertisements. Emotion recognition can also significantly improve Human-Computer Interaction to provide intuitive responses to the emotional states of users while using virtual assistants, video conferencing, or even during customer service interactions.

To develop such a system, we utilized Deep Learning as deep learning models are extremely efficient in recognizing images. They excel at extracting complex patterns and features from images, allowing us to accurately detect and classify emotions. Specifically, YOLO is effective for accurate face detection in real time, while CNNs enable the extrapolation of intricate patterns and features for each class of facial expressions corresponding to distinct emotions. Further, deep learning is crucial to generalizing emotions since these models will be trained on large datasets of facial expressions such that the system continuously improves in classifying different emotions. This highlights the effectiveness of deep learning within the context of our model.

2 ILLUSTRATION / FIGURE

An illustration of the model highlighting the input and expected output of the model for the given instance is depicted in Figure 1.

3 BACKGROUND & RELATED WORK

The following initiatives detail the work carried out in the field of real-time facial expression recognition through deep learning:

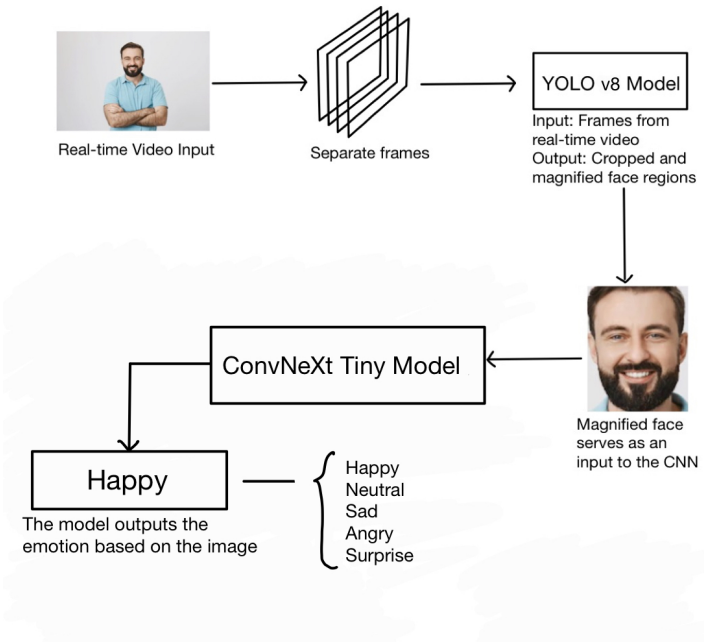

Figure 1: Integrated Model Illustration

3.1 IMAGE-BASED FACIAL EMOTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK ON EMOGNITION DATASET (1)

This work focuses on utilizing deep learning neural networks for facial expression recognition (FER) based on the Emognition dataset for ten target emotions that include awe, liking, enthusiasm, amusement, surprise, disgust, anger, fear, neutral, and sadness. It details how an automated approach using Convolutional Neural Networks can enable diverse classification through end-to-end learning using extensive data, thus directly relating to our project. Additionally, the project receives a 96% accuracy on the testing data, demonstrating successful performance.

3.2 A HYBRID FACIAL EXPRESSION RECOGNITION SYSTEM BASED ON RECURRENT NEURAL NETWORK (2)

In this paper, the authors present a sequence-based robust feature extraction framework with Recurrent Neural Networks for facial expression recognition. The system's performance is evaluated using the CK+ and the Oulu-CASIA dataset both of which have 6 emotion labels in common, namely anger, disgust, fear, happiness, sadness and surprise. In addition, the hybrid RNN model achieves an average testing accuracy of 95%, comparable to the CNN approach.

3.3 AN EFFICIENT DEEP LEARNING TECHNIQUE FOR FACIAL EMOTION RECOGNITION (3)

This study presents a deep learning model for real-time emotion, gender, and age classification using CNNs. Trained on JAFFE, CK+, and UTKFACE datasets, it achieved 95.65% accuracy for emotions and 98% for age and gender classification. It serves as a benchmark for comparing real-time performance with our emotion recognition model.

3.4 A GRAPH-STRUCTURED REPRESENTATION WITH BRNN FOR STATIC-BASED FACIAL EXPRESSION RECOGNITION (4)

A new technique for identifying facial expressions in still photos is presented in this work. In order to recognize static expressions, it models the spatial and geometric relationships between facial

landmarks using Bidirectional Recurrent Neural Networks with a graph-based structure. To infer expressions, the final representation is run through Softmax and fully connected layers. According to experimental results, the model performs at a state-of-the-art level, achieving 93.06% accuracy on the Oulu-CASIA Database, 98.27% on CK+, and 94.44% on MMI.

3.5 REAL TIME EMOTION RECOGNITION FROM FACIAL EXPRESSIONS USING CNN ARCHITECTURE (5)

This work proposes a Convolutional Neural Network based LeNet architecture for detecting emotions from facial expressions in real time. It presents a low cost and functionality model to classify seven different emotions, with the training and evaluation completed on a large amount of data gathered through the merging of 3 datasets: JAFFE, KDEF, and a custom dataset. Additionally, the model achieves a favorable validation and testing accuracy through the use of the custom database.

4 DATA PROCESSING

For data processing, we are using two publicly available Kaggle datasets: the Face Expression Recognition Dataset by jonathanoheix (6) and the Random Images for Face Emotion Recognition by sudarshanvaidya (7). We combined these datasets to create our own dataset for five emotions: anger, neutral, happy, sad, and surprise. Then we applied some data cleaning steps to ensure the quality of the dataset. We manually removed any irrelevant images, as well as those that did not correspond to the specified target emotions. We also removed images that were not properly aligned or were distortion, which could negatively impact model performance by introducing noise.

Additionally, we resized all images to a uniform size of 48x48 pixels and normalized pixel values to a range of [0, 1] by scaling down the RGB values, ensuring consistency across inputs and improving model stability during training. The cleaned and standardized dataset contains the following distribution of images:

- Anger: 4030
- Neutral: 4120
- Happy: 4281
- Sad: 4077
- Surprise: 4002

Total: 20510

Figure 2: The figure shows samples for each target emotion: (a) neutral, (b) happy, (c) sad, (d) angry, and (e) surprise

To enhance the diversity of the training dataset, we applied a series of augmentations (as shown in Figure 3) to introduce variability and improve generalization. We randomly applied horizontal flipping and random rotations up to 15 degrees, allowing the model to handle slight misalignments. We also applied color augmentations using ColorJitter, where brightness, contrast, and saturation were set to 30% and hue was set to 0.1. This modification makes the model less sensitive to lighting changes. Affine transformations with a translation factor of up to 10% of the image dimensions were applied to introduce minor shifts, mimicking slight variations in positioning within the frame. We also added a random distortion with a 20% distortion factor, helping the model generalize to cases where faces may not be perfectly clear. These augmentations help to simulate real-world variations and improve the model's ability to generalize effectively to new, unseen data.

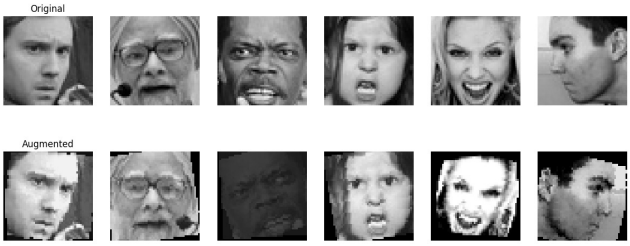

Figure 3: Images with various transformations applied to them at random

We used a 70-15-15 split for training, validation, and testing. This division provides a large and balanced dataset for model learning while keeping sufficient data aside for unbiased testing on unseen data. Each class is proportionally represented in all three subsets, ensuring that no emotion is over- or under-represented at any stage of model development.

For real-time performance evaluation, we created a separate set of videos with individuals naturally displaying various emotions. These videos are not part of the original dataset and serve as a completely unseen dataset, simulating real-world applications. This approach allows us to assess the model's real-world performance.

One challenge we faced was managing inconsistencies between datasets, as they included some unrelated images. We spent additional time on manually removing unrelated data points to ensure all samples were relevant. Also, fear and disgust were creating confusion in the model due to their similarity and low data count, which would often result in misclassification. We ultimately decided to remove these to classes to improve the model's overall performance.

5 Architecture

5.1 Description

The model architecture is divided into two particular sections - YOLO and CNN. Since the CNN is specifically trained on static face images to categorize emotions, we require the YOLO model to only pass magnified face images to the CNN for accurate emotion recognition. Further, the CNN is the most suitable neural network to categorize emotions based on facial images due to its ability to capture intricate patterns in the provided images.

5.2 Model Architecture - YOLO

The YOLO (You Only Look Once) model, the face detection component that distinguishes and captures facial regions in real-time video frames, is a major component of this emotion recognition model. By constructing bounding boxes with a confidence score around identified regions, YOLO, a highly trained face identification system, efficiently recognizes faces by scanning every frame in a single pass. These bounded portions are cropped to create zoomed-in images of the faces, and the CNN is then used to categorize the emotions. By focusing on the facial regions, YOLO lowers background noise, streamlines CNN data input, and increases the model's overall accuracy and speed in identifying emotions from facial expressions. The system can precisely and instantly identify emotions thanks to this technique.

To take advantage of the YOLO model's strong face identification capabilities without requiring a lot of retraining, we use a pre-trained version of the model.

5.3 Model Architecture - CNN

The architecture of the primary model, specifically the CNN, is based on the ConvNeXt Tiny neural network, which is well-suited for image classification tasks. We have further tuned the hyperparameters, and applied regularization and normalization to minimize training and validation loss and improve the testing accuracy of the model. The features of the model are as follows (8):

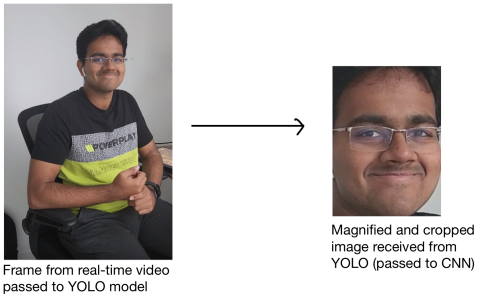

Figure 4: Function of YOLO Model

- The initial block of ConvNeXt Tiny consists of a large convolution, specifically with a 4x4 kernel with stride 4, to downsample the input image, along with the application of layer normalization.
- ConvNeXt Blocks make up each of the four stages that make up ConvNeXt Tiny. Figure 5 displays the number of channels and blocks for each stage. For stability and quick convergence, depthwise convolution with layer normalization is used in each step. Non-linearity is added by a GELU activation and a 1x1 convolution. To improve feature extraction, the stages increase output channels and downsample feature maps with a stride of 2.
- Global Average Pooling is also applied at the final stage, which reduces each channel to a single value.
- At the end, the output from the final stage and the pooling is passed to a fully connected layer which outputs 5 neurons for our 5 classes - Angry, Happy, Sad, Neutral, and Surprise.

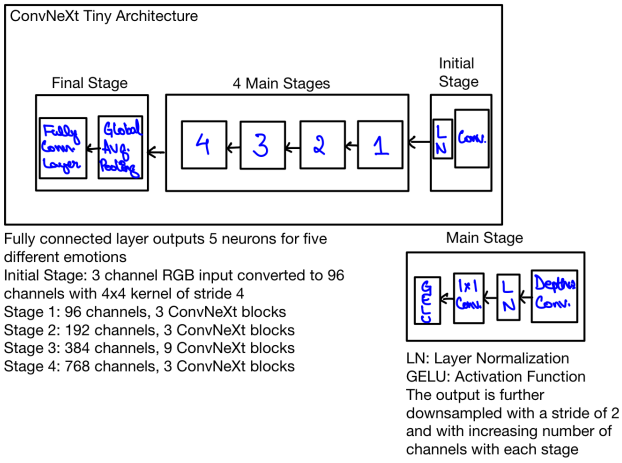

Figure 5: CNN Architecture

Thus, by leveraging transfer learning through the ConvNeXt Tiny neural network, we were able to use pre-trained weights, which reduced the training time significantly and improved the performance of our model. As for the hyperparameters, we explored a range of hyperparameter configurations to maximize accuracy and minimize loss, while preventing the model from overfitting. The best configuration for the hyperparameters, which will be ratified by our quantitative results, is as follows:

- Hyperparameters:
  - Learning Rate: 0.0001
  - Batch Size: 512

- Number of Epochs: 18
- Weight Decay: 1e-05

5.4 Baseline Model

To establish a benchmark for our emotion recognition system, we first tested a Random Forest classifier. Despite its simplicity, it yielded poor accuracy, as shown in the image. We then tried a Support Vector Machine (SVM), which performed even worse, achieving only 41.08% accuracy and struggling with 'neutral' and 'sad' expressions. Both models benefited slightly from feature extraction methods like HOG, SIFT, and PCA but remained inadequate for the task's complexity. These results highlight the limitations of traditional models and justify the need for a more advanced CNN-based approach to handle nuanced emotion classification effectively.

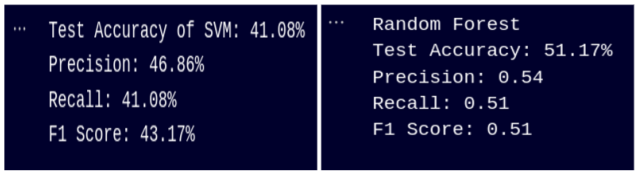

Figure 6: Baseline Model Metrics - SVM & Random Forest

6 Quantitative Results

Our emotion recognition model's quantitative performance was assessed using multiple metrics in order to fully demonstrate both its strengths and weaknesses. We first established a baseline using a Support Vector Machine (SVM) model, which achieved an accuracy of 41.08%. This result emphasises how difficult it is to classify emotions using simpler algorithms, especially when handling complex features like facial expressions. On the other hand, our CNN-based model showed a notable improvement in performance by utilising ConvNeXt Tiny with transfer learning. We trained our model on various hyperparameters and finally achieved a training accuracy of 80.00%, validation accuracy of 79.31%, and testing accuracy of 78.27%. The close alignment between training and validation accuracy demonstrates effective generalization and little overfitting. We also evaluated

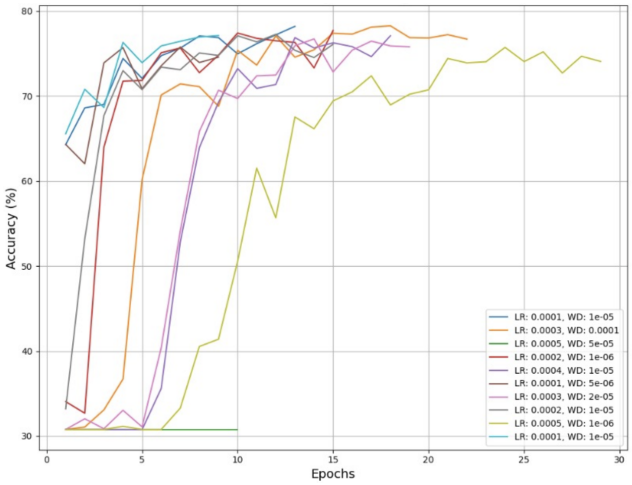

Figure 7: Training accuracy of various hyperparameters

the model's precision, recall, and F1 score, which were 0.8053, 0.7820, 0.7909 respectively. Importantly, the consistency between these metrics points to consistent performance across all emotion

classes, suggesting that the model is neither overfitting nor underfitting and is not biased towards any one emotion.

The improvement over the SVM baseline model highlights the effectiveness of our approach. These results demonstrate the model's potential for practical applications, such as customer interaction systems that enhance engagement by understanding user emotions. Overall, the model's strong and reliable performance establishes it as a promising solution for emotion recognition tasks.

7 Qualitative Results

The model's primary function is to classify various facial emotions based on expressions, with predictions effectively showcasing its performance. The model correctly identifies the exprections correctly in 78% of cases, as reflected in the testing accuracy.

As shown in figure 8 & figure 9, the model excels at recognizing distinct emotions such as "Happy" and "Surprise", as evident in the majority of predictions. However, challenges arise with ambiguous cases, such as neutral expressions with slight frowns, which are sometimes misclassified as "Sad."

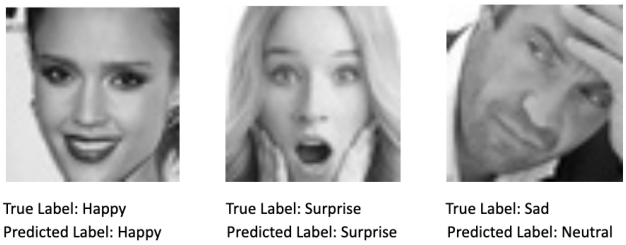

Figure 8: Predictions by model

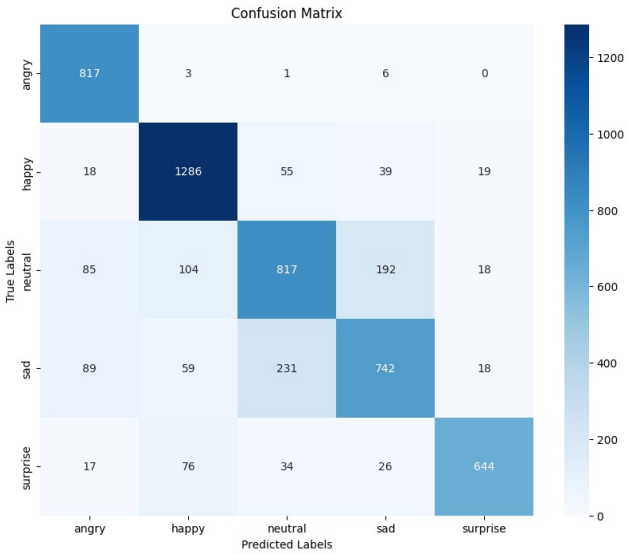

Figure 9: Confusion matrix of the model

8 Evaluation on New Data

To make sure the model's performance was completely and precisely assessed, we tested it using brand-new, untested data that had no influence at all on the model's design or hyperparameter tuning. This enabled us to evaluate its reliability and generalization ability in practical settings.

8.1 Testing Approach

8.1.1 Personalized Video Data:

- We created a video featuring one of our teammates displaying various facial expressions corresponding to the target emotions: Angry, Happy, Sad, Neutral, and Surprise.
- This video was processed using our model's pipeline, involving YOLO for face detection and the ConvNeXt Tiny CNN for emotion classification.
- Results were monitored in real time, and predictions were recorded.

8.2 Results

The model demonstrated strong performance on the new data, meeting expectations for the emotion recognition task:

- Predictions aligned well with the actual emotions displayed in the teammate's video as shown in figure 10 .

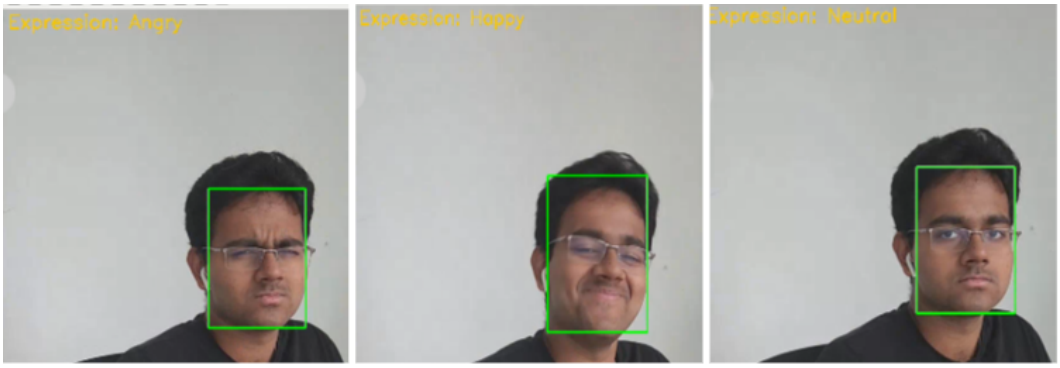

Figure 10: Predictions by model on real data

- Minimal latency was observed, showcasing the model's real-time processing capability.

8.3 Conclusion

Complete and objective efforts were made to test the model on fresh data. In the majority of cases, the outcomes surpassed expectations, demonstrating the model's resilience and suitability for use in practical situations. Future updates will address the minor misunderstanding between the emotions of sadness and neutrality. This assessment shows how well the model generalizes to novel and untested data.

9 Discussion

The model achieved a testing accuracy of 78.27%, reflecting reasonable performance for an emotion recognition task while highlighting opportunities for improvement. It excelled at identifying emotions with distinct facial expressions, such as happiness, anger, and neutrality. However, it didn't do as well while classifying subtler emotions, such as sadness, and faced even greater challenges with emotions like fear and disgust, which led to their exclusion from the final classifier due to insufficient representation in the dataset.

Another challenge was the model's difficulty in handling dynamic scenes, such as sharp transitions between emotions, as it struggled to capture the temporal context effectively. Additionally, the dataset's imbalance, with emotions like happiness and anger appearing more frequently, skewed predictions and increased confidence in these classes, even when incorrect.

To enhance performance, future work could integrate speech components for multimodal emotion recognition, combining visual and auditory cues to better interpret context-dependent emotions. Expanding the dataset to include underrepresented emotions such as sadness, fear, and disgust could also improve the classifier's ability to identify these categories with greater accuracy.

Overall, the team has done an extremely commendable job in obtaining a 78.27% accuracy, along with a 80% training accuracy, indicating an effective performance across various emotions with a low rate of misclassification.

10 Ethical Considerations

The system can be misused in a number of ways and hence careful use of the system must be ensured.

10.1 Privacy Violations and Prejudice

The system could be used for unauthorized emotion tracking in settings without consent, raising privacy concerns. Organizations could deploy the system in workplaces or public spaces to monitor employees or individuals and may violate the right to privacy. Storing or sharing this data could expose sensitive information to outsiders, leading to ethical implications. Biased training data could cause the model to misinterpret emotions based on cultural or demographic factors, instigating discrimination.

10.2 Model and Data Limitations

The model focuses solely on facial expressions and does not account for the situational context of emotions, leading to potential misinterpretations. The model may have some bias and prejudice.
The training data has a large amount of "angry", "happy" and "neutral" images, and hence, the classification of these emotions are very good. However, subtle emotions such as sadness and surprise have fewer images, and hence is not easy for the system to classify. Furthermore, the model struggles to distinguish subtle differences between anger and sadness.

11 Project Difficulty and Quality

Emotion recognition is extremely unpredictable due to the subtle differences in emotions such as 'sad' and 'neutral', variability in facial expressions across individuals, and even dataset limitations. In addition, limitations presented by GPU hardware significantly expand training time, resulting in a tedious process to ensure that the model performs as desired. Despite these hurdles, our model performs extremely well achieving a high testing accuracy, going beyond the foundational material provided through the integration of a YOLO-based face detection model along with a precisely tuned ConvNext Tiny CNN model. To address imbalances in data, we used data augmentation strategies such as rotations, contrast and brightness adjustments to ensure diversity within the training samples. In addition, we also parsed pre-recorded videos through our model while providing real-time emotion recognition.

Ultimately, our results highlight how a complex and inherently difficult problem such as real-time emotion recognition can be effectively addressed using the right design choices and project decisions. With more computational power, we could have trained a deeper neural network within the same system to achieve a greater testing accuracy.

References

[1] Erlangga Satrio Agung, Achmad Pratama Rifai, and Titis Wijayanto. Image-based facial emotion recognition using convolutional neural network on emognition dataset, Jun 2024.

[2] Jing-Ming Guo, Po-Cheng Huang, and Li-Ying Chang. A hybrid facial expression recognition system based on recurrent neural network — ieee conference publication — ieee xplore, Nov 2019.

[3] Asad Khattak, Muhammad Zubair Asghar, Mushtaq Ali, and Ulfat Batool. An efficient deep learning technique for facial emotion recognition - multimedia tools and applications, Oct 2021.

[4] Lei Zhong, Changmin Bai, Jianfeng Li, Tong Chen, Shigang Li, and Yiguang Liu. A graph-structured representation with brnn for static-based facial expression recognition — ieee conference publication — ieee xplore.

[5] Mehmet Akif Ozdemir, Berkay Elagoz, Aysegul Alaybeyoglu, Reza Sadighzadeh, and Aydin Akan. Real time emotion recognition from facial expressions using cnn architecture — ieee conference publication — ieee xplore, Nov 2019.

[6] Jonathan Oheix. Face expression recognition dataset, Jan 2019.

[7] Sudarshan Vaidya. Natural human face images for emotion recognition, Dec 2020.

[8] Zhuang Liu and Hanzi Mao. A convnet for the 2020s.

(6) (7) (8)