ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 5 May 2024

Xface: Facial Detection and Emotion Profiling

[1] P. Murali*, [2] Deepti Busennagari, [3] Aditya Rawat

[1] [2] [3] Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur 603203
Corresponding Author Email: [1] muralip@srmist.edu.in

Abstract*— Emotion recognition is a crucial aspect of human interaction, yet individuals with alexithymia face challenges in understanding and expressing emotions. This paper proposes a Realtime Emotion Analysis system using facial recognition technology to assist individuals with alexithymia in interpreting facial expressions. Leveraging advanced face detection algorithms such as Multi-task Cascaded Convolutional Networks (MTCNN) and visual transformers for emotion recognition, the system aims to provide accurate real-time feedback on facial expressions, thereby enhancing social interactions for neurodiverse individuals. Through the integration of diverse datasets and rigorous evaluation of performance metrics, the proposed system demonstrates promising results in supporting individuals with alexithymia in navigating social interactions effectively.

***Index Terms*— Alexithymia, Facial Recognition, Realtime Emotion Analysis, Facial Expression Recognition, Neurodiversity.**

I. INTRODUCTION

Emotion is a universal language that plays a fundamental role in human communication and interaction. From conveying joy and excitement to expressing sadness and anger, emotions shape our perceptions, behaviors, and relationships [6]. However, for individuals with alexithymia, understanding and expressing emotions can be a significant challenge. Alexithymia, characterized by difficulties in identifying and describing one's emotions, affects approximately 8% of males and 2% of females, according to research studies [7]. This neuropsychological phenomenon not only impacts an individual's personal well-being but also hinders their ability to navigate social interactions effectively.

Recognizing the importance of addressing the needs of individuals with alexithymia, researchers and technologists have explored innovative solutions to support emotion recognition and understanding. Facial recognition technology has emerged as a promising avenue for assisting individuals with alexithymia in interpreting facial expressions, a key aspect of emotional communication. By leveraging advanced algorithms and machine learning techniques, facial recognition systems can analyze facial features and gestures to infer underlying emotional states.

One of the key components of facial recognition systems is face detection, which involves identifying and locating faces within images or video streams. Multi-task Cascaded Convolutional Networks (MTCNN) is a popular algorithm used for face detection, known for its accuracy and efficiency. MTCNN employs a multi-stage approach, including Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net), to detect facial regions with varying scales and orientations [1][2]. By accurately localizing faces, MTCNN lays the foundation for subsequent facial feature extraction and emotion recognition processes.

Facial feature extraction involves identifying key points and landmarks on detected faces, which serve as input to the emotion recognition module. In recent years, visual transformers have emerged as a powerful architecture for facial expression recognition tasks [8]. Inspired by the Transformer architecture initially developed for natural language processing, visual transformers utilize self-attention mechanisms to capture global dependencies within images [15]. This enables visual transformers to effectively model spatial relationships and extract relevant features from facial images, making them well-suited for emotion recognition tasks.

To train visual transformer models for facial expression recognition, diverse datasets containing annotated facial expressions are utilized. These datasets, such as CK+, FER2013, and facial emotion recognition dataset, provide a wide range of

[18] facial expressions, including happiness, sadness, anger, fear, surprise, and disgust. By fine-tuning pre-trained transformer models on these datasets, researchers can improve the model's accuracy and generalization capabilities, enabling it to recognize subtle emotional cues and nuances.

Once trained, the Realtime Emotion Analysis system can provide real-time feedback on recognized facial expressions, facilitating improved social interactions for individuals with alexithymia [7]. Performance metrics, including accuracy, precision, recall, F1-score, false positives, false negatives, true positives, true negatives, and processing speed, are calculated to evaluate the effectiveness of the system. Additionally, the system's performance is compared with existing approaches and benchmark datasets to assess its accuracy and efficiency comprehensively. Through continued research and development, facial recognition technology holds the potential to significantly improve the quality of life for individuals with alexithymia, fostering greater understanding and empathy in social interactions.

II. RELATED WORK

The development of facial recognition technology and emotion analysis systems has been a topic of extensive research in recent years, driven by the increasing demand for innovative solutions to support individuals with neurodiverse conditions such as alexithymia. In this section, we review relevant studies and advancements in the field of facial recognition, emotion recognition, and technology-based interventions for individuals with alexithymia.

The cascade face detector proposed by Viola and Jones [5] utilizes Haar-Like features and AdaBoost to train cascaded classifiers, which achieve good performance with real-time efficiency. However, quite a few works [1, 3] indicate that this detector may degrade significantly in real-world applications with larger visual variations of human faces even with more advanced features and classifiers. Zhang et al. (2016) proposed MTCNN, a multi-stage face detection framework that achieves state-of-the-art performance on benchmark datasets [2]. The multi-stage architecture of MTCNN enables robust detection of faces across different scales and orientations, making it suitable for real-world applications [15].

In addition to face detection, facial feature extraction and emotion recognition play crucial roles in facial recognition systems. Recent studies have explored the use of deep learning techniques, such as convolutional neural networks (CNNs) and visual transformers, for facial expression recognition. Liu et al. (2020) [14] proposed a visual transformer-based approach for facial expression recognition, demonstrating superior performance compared to traditional CNN-based methods. Visual transformers leverage self-attention mechanisms to capture global dependencies within facial images, enabling accurate recognition of subtle emotional cues.

Moreover, the availability of large-scale annotated datasets, such as CK+, FER2013, and facial emotion recognition dataset, has facilitated the development and evaluation of emotion recognition systems. These datasets contain diverse facial expressions captured under various conditions, providing researchers with valuable resources for training and testing their models. A comprehensive review of benchmark datasets and evaluation metrics is essential for benchmarking the performance of emotion recognition systems accurately.

Emotion Recognition Systems: Emotion recognition systems aim to infer emotional states from facial expressions, gestures, and vocal cues. Traditional approaches to emotion recognition relied on handcrafted features and machine learning classifiers. However, recent advancements in deep learning have revolutionized the field, enabling end-to-end learning of complex patterns from raw data. Zhang et al. (2018) [16] proposed an end-to-end deep learning framework for multimodal emotion recognition, combining visual and auditory cues to improve accuracy.

Additionally, researchers have explored the use of recurrent neural networks (RNNs) and attention mechanisms for sequential emotion recognition tasks. Zhao et al. (2016) [10] introduced the Attention-Based Recurrent Neural Network (AB-RNN) for facial expression recognition in videos, achieving state-of-the-art performance on benchmark datasets such as the FDDB dataset [12]. The AB-RNN model leverages attention mechanisms to focus on relevant facial regions over time, capturing temporal dependencies in facial expressions.

Furthermore, the integration of multimodal information, such as facial expressions, body language, and speech, has shown promise in enhancing emotion recognition accuracy. Fusion strategies, such as late fusion and early fusion, enable the combination of complementary modalities to improve overall performance. Zhao et al. (2019) [15] proposed a multimodal emotion recognition framework that combines facial expressions and speech features, achieving superior performance compared to unimodal approaches.

Technology-Based Interventions for Alexithymia: Technology-based interventions have shown promise in supporting individuals with alexithymia in understanding and expressing emotions. Virtual reality (VR) environments, for example, offer a controlled and immersive platform for practicing social interactions and emotional expression. Kau et al. (2021) developed a VR-based intervention program for individuals with alexithymia, incorporating emotion recognition tasks and social scenarios to improve emotional awareness and communication skills.

Moreover, mobile applications and wearable devices hold potential as tools for assisting individuals with alexithymia in real-world settings. Emotion tracking apps, such as Mood Notes and Mood Meter, enable users to monitor their emotional states and track patterns over time. Wearable sensors, equipped with physiological monitoring capabilities, can provide real-time feedback on stress levels and emotional arousal, empowering individuals with alexithymia to manage their emotions more effectively.

Conclusion: In summary, the development of facial recognition technology and emotion analysis systems has opened new avenues for supporting individuals with alexithymia in understanding and expressing emotions. Leveraging advanced algorithms such as MTCNN and visual transformers, along with large-scale annotated datasets, enables the creation of accurate and efficient Realtime Emotion Analysis systems. Moreover, technology-based interventions, including VR environments, mobile applications, and wearable devices, offer personalized solutions for improving emotional awareness and communication skills in individuals with alexithymia [9]. Through continued research and innovation, technology has the potential to significantly enhance the quality of life for neurodiverse populations, fostering empathy, understanding, and inclusion in society.

III. III. BACKGROUND

Alexithymia is a neuropsychological phenomenon characterized by difficulties in recognizing, expressing, and describing one's emotions. It is often associated with a range of psychological disorders, including autism, depression, and schizophrenia. Individuals with alexithymia may struggle with both the cognitive and affective dimensions of emotional experience, leading to challenges in interpersonal relationships and mental well-being. Additionally, face blindness, or prosopagnosia, presents further obstacles in recognizing and remembering faces, exacerbating social difficulties for affected individuals. Addressing the needs of neurodiverse populations requires innovative solutions that leverage advances in technology and artificial intelligence.

Challenges and Limitations in Existing Systems:

Existing face detection systems face several challenges, including scale variability, occlusion, pose variability, illumination conditions, complex backgrounds, limited training data, and computational resource requirements. These challenges pose significant obstacles in accurately detecting and recognizing facial expressions, particularly in real-time scenarios. Overcoming these limitations is crucial for developing an effective Realtime Emotion Analysis system that can support individuals with alexithymia and other neurodiverse conditions.

IV. PREDICTION MODEL CONSTRUCTION

The development of a prediction model for Realtime Emotion Analysis using Facial Recognition involves a systematic approach encompassing data preprocessing, model selection, training, evaluation, and integration stages. Leveraging advanced algorithms such as Multi-task Cascaded Convolutional Networks (MTCNN) for face detection and visual transformers for emotion recognition, the construction of the prediction model aims to provide accurate and reliable analysis of facial expressions in real-time. The following outlines the key components and methodologies involved in the construction of the prediction model:

A. Data Gathering and Preprocessing

The first step in constructing the prediction model involves gathering and preprocessing the dataset. This entails compiling a diverse dataset containing annotated facial expressions captured under various conditions. Additionally, relevant metadata such as timestamps, environmental factors, and contextual information may be incorporated to enhance the understanding of facial expressions. Preprocessing techniques such as data cleaning, normalization, and feature extraction are applied to ensure data consistency and quality. This includes handling missing values, removing outliers, and standardizing the data to facilitate model training.

B. Machine Learning Models

The selection of appropriate algorithms and models is crucial in constructing an effective prediction model for Realtime Emotion Analysis. Given the complexity of facial expression recognition tasks, a combination of state-of-the-art techniques is employed. Multi-task Cascaded Convolutional Networks (MTCNN) are utilized for face detection, owing to their robustness and efficiency in detecting faces across different scales and orientations. Additionally, visual transformers are employed for emotion recognition, leveraging self-attention mechanisms to capture global dependencies within facial images. The integration of these models allows for comprehensive analysis of facial expressions and enhances the prediction accuracy.

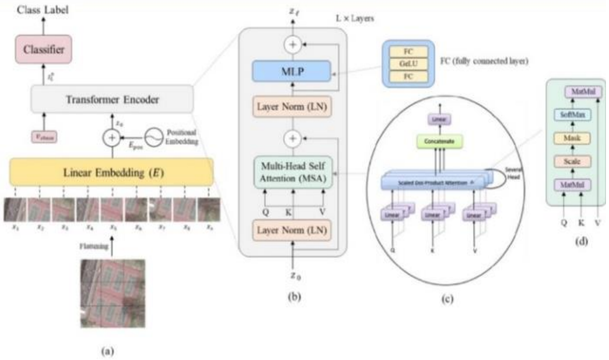


Figure 1. Virtual Transformers Architecture

C. Model Training

Once the models are selected, they are trained using the preprocessed dataset. The training process involves optimizing model parameters to minimize the prediction error and maximize accuracy. For MTCNN, the model is trained using annotated facial images to learn the patterns and features indicative of facial regions. Similarly, visual transformers are trained on the dataset to recognize and classify different facial expressions. Techniques such as data augmentation, regularization, and cross-validation are employed to prevent overfitting and improve generalization performance.

D. Evaluation

After training, the prediction model is evaluated using appropriate performance metrics to assess its effectiveness and reliability. Common metrics include accuracy, precision, recall, F1-score, and confusion matrix analysis. The model is tested on a separate validation dataset to measure its ability to generalize to unseen data. Evaluation results provide insights into the strengths and limitations of the prediction model, guiding further refinement and optimization efforts.

E. Integration

Once the prediction model is trained and evaluated, it is integrated into the Realtime Emotion Analysis system architecture. This involves developing interfaces and APIs to facilitate seamless communication between the prediction model and other system components. The integration ensures that the prediction model can effectively analyze facial expressions in real-time and provide timely feedback to users. Additionally, ongoing monitoring and maintenance of the prediction model are essential to ensure its continued accuracy and effectiveness in real-world scenarios.

In conclusion, the construction of a prediction model for Realtime Emotion Analysis using Facial Recognition involves a structured approach encompassing data preprocessing, model selection, training, evaluation, and integration stages. By leveraging advanced algorithms such as MTCNN and visual transformers, the prediction model enables accurate and reliable analysis of facial expressions in real-time, contributing to enhanced social interactions and emotional understanding for individuals with alexithymia.

V. METHODOLOGY

A. Data Collection and Preprocessing

The first step in the methodology involves gathering a diverse dataset of facial images annotated with corresponding emotional labels. This dataset comprises expressions of various intensities, captured under different lighting conditions, facial orientations, and backgrounds. Additionally, metadata such as timestamps, environmental factors, and contextual information may be collected to provide further insights into the emotional context.

Once collected, the dataset undergoes preprocessing to ensure data consistency and quality. This includes data cleaning to handle missing values, removal of outliers, and normalization of pixel values to a standardized

B. Face Detection

The next stage involves face detection, where Multi-task Cascaded Convolutional Networks (MTCNN) are utilized to identify and localize faces within the images. MTCNN employs a multi-stage architecture comprising three stages: Proposal Network (P-Net), Refinement Network (R-Net), and Output Network (O-Net) [3]. This multi-stage approach enables robust detection of faces across different scales and orientations, ensuring accurate localization even in challenging conditions [11].

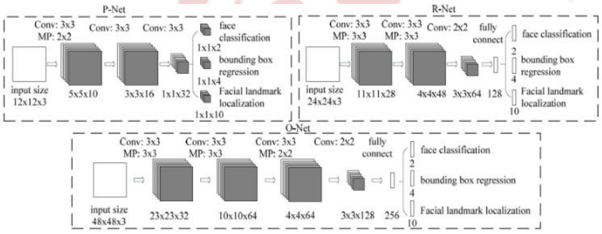


Figure 2. MTCNN Architecture

C. Facial Expression Recognition

Following face detection, the detected faces are passed through visual transformers for facial expression recognition. Visual transformers leverage self-attention mechanisms to capture global dependencies within facial images, enabling accurate recognition of subtle emotional cues. The transformer architecture consists of multiple layers of self-attention and feedforward neural networks, allowing the model to effectively extract features and classify facial expressions.

The visual transformer model is trained on the preprocessed dataset using supervised learning techniques. During training, the model learns to map input facial images to corresponding emotional labels through backpropagation and gradient descent optimization. Hyperparameter tuning, regularization techniques, and data augmentation are employed to prevent overfitting and improve model generalization.

D. Model Integration:

Once trained, the face detection components are integrated into a unified Realtime Emotion Analysis system architecture. This integration involves developing interfaces and APIs to facilitate seamless communication between the different components of the system. Real-time processing capabilities are implemented to ensure timely analysis and feedback on detected facial expressions.

E. Evaluation and Validation

The performance of the Realtime Emotion Analysis system is evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. The system is tested on a separate validation dataset to assess its ability to generalize to unseen data. Additionally, qualitative evaluation through user studies and feedback is conducted to validate the system's effectiveness in real-world scenarios.

F. Optimization and Refinement:

Continuous optimization and refinement of the system are essential to enhance its accuracy, speed, and robustness. This involves fine-tuning model parameters, updating training data, and incorporating feedback from user testing. Additionally, advancements in facial recognition algorithms and techniques are monitored and integrated into the system to stay abreast of the latest developments in the field.

VI. PROPOSED SOLUTION:

The proposed solution incorporates the use of [15] visual transformers for facial expression recognition, implemented in PyTorch, to complement the [18] Multi-task Cascaded Convolutional Networks (MTCNN) for improved facial detection. Visual transformers have emerged as a powerful architecture for image processing tasks, leveraging the self-attention mechanism to capture long-range dependencies within images. By integrating visual transformers into the Realtime Emotion Analysis system, we can enhance the accuracy and robustness of both facial detection and expression recognition, particularly in complex social contexts where individuals with alexithymia may exhibit subtle facial expressions.

Visual transformers, inspired by the Transformer architecture initially developed for natural language processing tasks, have gained popularity in computer vision due to their ability to effectively model spatial relationships in images. Unlike traditional convolutional neural networks (CNNs), which rely on local receptive fields, visual transformers employ self-attention mechanisms to capture global dependencies across different regions of an image. This enables visual transformers to effectively capture complex patterns and relationships within facial images, making them well-suited for tasks such as facial expression recognition.

One of the key advantages of visual transformers is their ability to handle long-range dependencies within images, which is particularly important for facial expression recognition tasks where subtle facial cues may be distributed across different regions of the face. By attending to relevant regions of the face, visual transformers can effectively capture important features and patterns that contribute to accurate recognition of facial expressions. This is crucial for individuals with alexithymia, who may exhibit nuanced expressions that require careful analysis for accurate interpretation.

Moreover, visual transformers offer scalability and flexibility, allowing for the integration of large-scale pre-trained models such as ViT (Vision Transformer) [17] for facial expression recognition tasks. Pre-trained visual transformer models [5], trained on large-scale image datasets, can capture rich representations of facial features, enabling accurate recognition of facial expressions with minimal fine-tuning. By fine-tuning the pre-trained visual transformers on additional training data specifically curated to include diverse facial expressions, we can further improve the model's ability to recognize subtle emotional cues and nuances, thus enhancing its effectiveness for individuals with alexithymia.

In addition to their effectiveness in capturing spatial relationships within images, visual transformers offer interpretability, enabling insights into the underlying features contributing to facial expression recognition decisions. By visualizing attention maps generated by visual transformers, researchers and practitioners can gain valuable insights into the salient regions of the face that contribute to different facial expressions. This interpretability is essential for understanding the model's behavior and identifying potential biases or limitations, particularly in sensitive applications such as emotion analysis for neurodiverse populations.

Furthermore, visual transformers can facilitate multi-modal fusion, enabling the integration of additional modalities such as audio and text for more comprehensive emotion analysis. By combining visual information from facial images with contextual cues from other modalities [15], the Realtime Emotion Analysis system can provide richer insights into individuals' emotional states, enhancing its utility for individuals with alexithymia and other neurodiverse conditions [8].

In summary, the integration of visual transformers into the Realtime Emotion Analysis system represents a significant advancement in facial expression recognition technology, particularly for individuals with alexithymia. By leveraging the self-attention mechanism and scalability of visual transformers, we can enhance the accuracy, interpretability, and multi-modal capabilities of the system, ultimately improving social interactions and quality of life for neurodiverse individuals. Through continued research and development in this area, we can unlock the full potential of visual transformers for emotion analysis and pave the way for more inclusive and supportive technologies in the future.

Core Algorithms for MTCNN

1) Loss function

Loss function is used to find the difference between the correct output and the predicted output of face features from a convolutional neural network model. The cross-entropy loss function is computed as follows:

$$L_i^{det} = -(y_i^{det} \log(p_i) + (1 - y_i^{det})(1 - \log(p_i)))$$

where p_i is the probability that the pic classified is a face, and det y_i denotes that the picture to be classified is a true and accurate label with a value range of {0,1}.

2) Intersection Over Union

IoU is a crucial metric for assessing segmentation models, commonly called Jaccard's Index, since it quantifies how well the model can distinguish objects from their backgrounds in an image The cross-to-match ratio is computed as follows:

$$IoU = \frac{S_{\text{Detection Result}} \cap S_{\text{Ground Truth}}}{S_{\text{Detection Result}} \cup S_{\text{Ground Truth}}}$$

where $S_{\text{Detection Result}}$ is the prediction frame and $S_{\text{Ground Truth}}$ is the real frame.

3) Bounding box regression

The result is usually directly discarded when IoU is less than the set threshold. Localizing multiple objects in an image is mainly done by bounding boxes. The bounding box is predicted with a loss function that gives the error between the predicted and ground truth bounding box. For the face image, the regression loss is calculated using the Euclidean distance:

$$L_i^{box} = \| y_i^{box} - y_i^{box} \|_2^2$$

where y_i^{box} is the coordinate obtained by network prediction, and y_i^{box} is the coordinate value of the real and accurate sample x_i, namely the four points. Therefore, $y_i^{box} \in R^4$.

4) Regression loss function for landmark positioning

The Euclidean distance between the predicted location of the landmark and the actual landmark is calculated, and the distance is minimized. For the face image x_i, the regression loss is calculated using the Euclidean distance:

$$L_i^{landmark} = \| y_i^{landmark} - y_i^{landmark} \|_2^2$$

where $y_i^{landmark}$ is the landmark location obtained by network prediction, and $y_i^{landmark}$ represents the actual real landmark coordinates. Since five points exist, each point has two coordinates. Therefore, $y_i^{landmark} \in R^{10}$.

5) Non – Maximum Suppression

Non-maximum suppression is a technique that is used after the region proposal step to eliminate duplicate bounding boxes and select the most relevant ones. The idea behind NMS is straightforward. It works by comparing the confidence scores of the proposed bounding boxes and eliminating the ones that overlap significantly with a higher-scoring bounding box. therefore, the prediction scores are relatively high. Finally, the tall candidate frame will be retained.

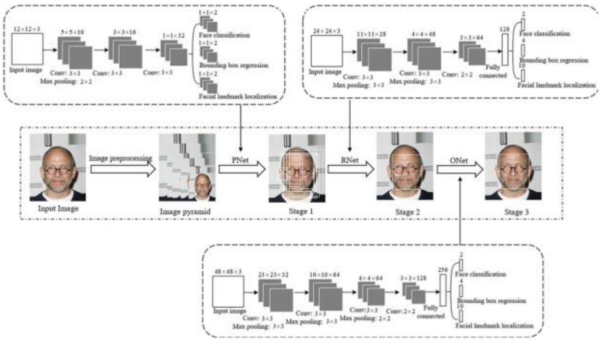


Figure 3. MTCNN Network

VII. RESULTS

[3] This paper introduces the important basic model MTCNN at first. It is one of the hottest models used most widely recently for its high precision and outstanding real time performance among the state-of-art algorithms for face detection. Then, the first basic application of portrait classification is researched based on MTCNN and FaceNet. Its direction is one of the most classical and popular area in nowadays AI visual research, and is also the base of many other industrial branches. In addition to the discussion to the basic models, several practical methods are also advised to improve the precision [4]. The FDDB dataset is used to compare with other algorithms [12]. The FER2013 dataset was used for ViT model, which was finetuned by preprocessing using the hugging face. We achieved an accuracy of 68.45%, which is more than the average 65% because of the large dataset. The highest recorded is 74%. Further improvements might need more computational ram.

One can improve the performance of the vision transformer models implemented in this paper using regularization and dropout techniques.

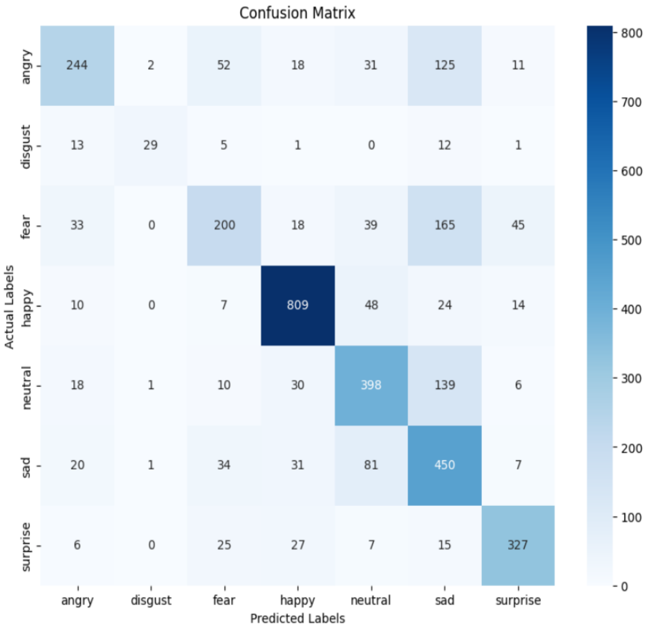


Figure 4. Confusion Matrix on FER2013

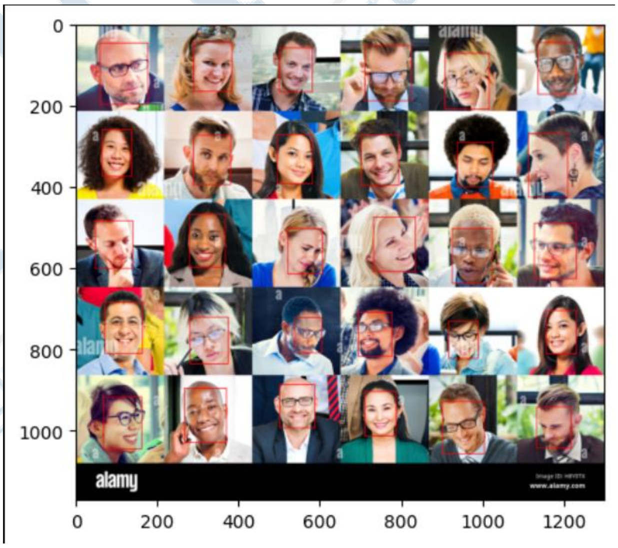
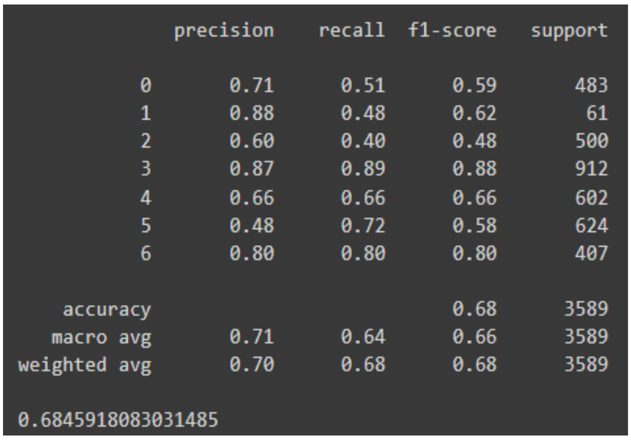


Figure 5. MTCNN Output



```
              precision    recall  f1-score   support

           0       0.71      0.51      0.59       483
           1       0.88      0.48      0.62        61
           2       0.60      0.40      0.48       500
           3       0.87      0.89      0.88       912
           4       0.66      0.66      0.66       602
           5       0.48      0.72      0.58       624
           6       0.80      0.80      0.80       407

    accuracy                           0.68      3589
   macro avg       0.71      0.64      0.66      3589
weighted avg       0.70      0.68      0.68      3589

0.6845918083031485
```

Figure 6. Emotion Analysis on FER2013 dataset

VIII. SCOPE AND APPLICATION

The proposed Realtime Emotion Analysis system has broad applications, including security and surveillance, human-computer interaction, biometric authentication, and integration with mental health apps. By integrating facial recognition technology with mental health apps, individuals with alexithymia and other neurodiverse conditions can receive personalized support and guidance in navigating social interactions. Furthermore, the system's ability to target a wide range of facial expressions enhances its applicability and effectiveness in real-world scenarios.

IX. CONCLUSION

In conclusion, the development of a Realtime Emotion Analysis system using facial recognition technology holds immense potential for supporting individuals with alexithymia and other neurodiverse conditions. By leveraging advanced algorithms such as MTCNN and CNNs, this system can provide real-time feedback on facial expressions [3], thereby enhancing social interactions and improving the quality of life for neurodiverse individuals. Moving forward, it is essential to address ethical considerations and conduct further research to validate the effectiveness and reliability of the proposed system in diverse real-world settings. Through interdisciplinary collaboration and technological innovation, we can strive towards creating a more inclusive and supportive society for all individuals, regardless of their neurodiversity.

REFERENCES

[1] Li H., Lin Z., Shen X., Brandt J. and Hua G. 2015 A convolutional neural network cascade for face detection IEEE Conference on Computer Vision and Pattern Recognition 5325-5334

[2] Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Senior Member, IEEE, and Yu Qiao, Senior Member, IEEE

[3] Juan Du 2020 J. Phys.: Conf. Ser. 1518 012066 High-Precision Portrait Classification Based on MTCNN and Its Application on Similarity Judgement

[4] Face Detection Based on Receptive Field Enhanced Multi-Task Cascaded Convolutional Neural Networks XIAOCHAO LI 1,2, (Senior Member, IEEE), ZHENJIE YANG 1, AND HONGWEI WU3, (Member, IEEE)

[5] P. Viola and M. J. Jones, “Robust real-time face detection. International journal of computer vision,” vol. 57, no. 2, pp. 137-154, 2004

[6] A. T. Kabakus, "PyFER: A Facial Expression Recognizer Based on Convolutional Neural Networks," in IEEE Access, vol. 8, pp. 142243-142249, 2020, doi: 10.1109/ACCESS.2020. 3012703.

[7] Facial Emotion Recognition Predicts Alexithymia Using Machine Learning Nima Farhoumandi, Sadegh Mollaey, Soomaayeh Heysieattalab, Mostafa Zarean, Reza Eyvazpour

[8] Xing Jin, Xulin Song, Xiyin Wu, Wenzhu Yan, " Transformer embedded spectral-based graph network for facial expression recognition ", International Journal of Machine Learning and Cybernetics, 2023.

[9] Emotion Analysis Based on Deep Learning with Application to Research on Development of Western Culture Front. Psychol., 13 September 2022 Sec. Emotion Science.

[10] Z. Zhao et al., "Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition," in IEEE Access, vol. 7, pp. 97515-97525, 2019, doi: 10.1109/ACCESS.2019. 2928625.

[11] Face Detection Based on Improved Multi-task Cascaded Convolutional Neural Networks Siyu Jia, Ying Tian IAENG International Journal of Computer Science

[12] V. Jain, and E. G. Learned-Miller, “FDDB: A benchmark for face detection in unconstrained settings,” Technical Report UMCS-2010-009, University of Massachusetts, Amherst, 2010.

[13] Howard, A.G, Zhu, M, Chen, B, Kalenichenko, D, Wang, W, Weyand, T, Andreetto, M, & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv, abs/1704.04861.

[14] Y. Liu et al., "A Survey of Visual Transformers," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2022.3227717.

[15] Facial Expression Recognition Based on Visual Transformers and Local Attention Features Network Shuang Zhao; Chang Liu; Guangyuan Liu.

[16] M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[17] C. -W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 2017, pp. 583-588, doi: 10.1109/ICME.2017.8019296.

[18] Thomas Kopalidis, Vassilios Solachidis,Nicholas Vretos, Petros Daras. "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets", information, 2024