A MULTIMODAL GAN FRAMEWORK WITH 3D CNN FOR THE EMOTION DETECTION IN AUTISM SPECTRUM DISORDER

REPORT

IT5712 - PROJECT I

Submitted by

Rahul Prasanth D (2020506070)

Bala Natesh R M(2020506018)

Sanjay G (2020506080)

Guided by

Dr. J.Dhalia Sweetlin

Associate Professor

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

MADRAS INSTITUTE OF TECHNOLOGY

ANNA UNIVERSITY: CHENNAI 600 044

AUG 2023

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects millions of children worldwide. One of the significant challenges faced by autistic children is the difficulty in expressing and understanding emotions, which can affect the effective communication and social interactions. Traditional methods of emotion recognition often fall short in addressing the unique needs and characteristics of these people.

Therefore, this project aims to create a specialized system that comprehensively identifies and understands emotions in autistic children, aiming to bridge the communication gap and provide essential insights into their emotional experiences which also helps the educators, therapists, and caregivers to take care of them.

II. PROBLEM STATEMENT:

The project addresses the challenge of understanding and effectively supporting emotional expression in autistic children. Autistic children often encounter difficulties in conveying their emotions. Our primary objectives include the development of a system that comprehensively identifies and understands the emotions exhibited by autistic children. This system will consider various modalities, including facial expressions, hand gestures, and other body movements, to provide a elaborate view of their emotional states. To address the issue of occluded faces, we will employ Generative Adversarial Networks (GANs) to ensure that even when parts of the face are hidden from view, the system can generate the complete face from the occluded faces. Additionally, we will implement landmark detection techniques and predict facial action units, combined with the analysis of body movements, to enhance the accuracy of emotion prediction. It will help to understand how autistic children experience emotions by using information about where and when emotions happen on their faces, along with specific facial expressions, to develop ways to support and help them more effectively.

III. LITERATURE SURVEY

Facial Action Unit Detection via Adaptive Attention and Relation

Shao et al. (IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 32, 2023)

The authors propose an adaptive attention regression network by integrating the advantages of local attention predefinition and global attention learning, which can capture both predefined dependencies by landmarks in strongly correlated regions and facial globally distributed dependencies in weakly correlated regions. To simultaneously reason the specific pattern of each AU, the inter-dependencies among AUs, as well as the temporal correlations, they also propose an adaptive spatio-temporal graph convolutional network. Extensive experiments on benchmark datasets show that the approach achieves comparable performance in both constrained scenarios and unconstrained scenarios, and can accurately learn the regional correlation distribution of each AU. The Adaptive Attention Regression (AAR) method achieved an average F1-frame score of approximately 63.8 on the BP4D benchmark. The AAR network was tested on input images with misalignment errors and occlusions. If input images are severely misaligned AAR fails to precisely capture AU Region Of Interests (ROIs). The AAR does not explicitly process misalignment errors, such as explicitly learning rotation-invariant and scale-invariant features.

Action unit classification for facial expression recognition using active learning and SVM

Yao et al. (Multimed Tools Appl 80, 24287–24301 (2021))

This paper proposes a combination of active learning and SVM for the extraction of AUs and facial expression classification. Active learning uses the existing model to acquire new knowledge by simulating the process of human learning. Based on continuously accumulated information, the existing model can be corrected to become more accurate. SVM was utilized to classify different AUs and ultimately map them to their corresponding facial expressions. Different facial expressions, regardless of being female or male, had different recognition rates ranging from 90% to 95% for females and from 83% to 95% for males. Of the seven facial expressions, five expressions (joy, sadness, anger, hate, and neutral) were recognized correctly. Regardless of gender in the samples, the hate and neutral facial expressions seem to be more difficult to recognize than the joy and surprise expressions.

Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?

Kumar et al. (EEE TRANSACTIONS ON AFFECTIVE COMPUTING DOI 2023)

AU intensity estimation and dimensional emotion recognition models are trained based on the temporal dynamics of 3D facial expressions. The quality of 3D Morphable Models (3DMM) based expression features on the datasets of valence-arousal estimation and AU intensity estimation are extensively evaluated. A simple bi-directional Gated Recurrent Unit (GRU) network is applied to model the temporal dynamics of 3DMM expression features extracted from five dense 3D face alignment models: ExpNet, 3DDFA-V2, RingNet, DECA and EMOCA. The recognition performance of different 3D face shape models with the 2D face appearance baselines and models are compared. In the case of continuous emotion recognition 3D face expression features outperform the existing benchmarks as well as the 2D appearance baselines. On the task of AU intensity prediction 3D face shape models perform poorly compared to the existing state-of-the-art benchmarks based on 2D appearance features. The MSE values for 2D and 3D shapes were found to be 0.36 and 0.53 respectively. The poor AU intensity estimation performance of the 3D face models might be due to the use of a global basis vector for expression modelling.

Multitask, Multilabel, and Multidomain Learning With Convolutional Networks for Emotion Recognition

Pons et al. (IEEE TRANSACTIONS ON CYBERNETICS, VOL. 52, NO. 6, JUNE 2022)

A datasetwise selective sigmoid cross-entropy loss function is formalized to simultaneously train a multitask, multilabel and multidomain model. The authors utilize two popular convolutional neural network (CNN) architectures, VGG-16 and Resnet-50, for their experiments. These networks serve as the basis for training models for emotion recognition and AU detection. Dedicated individual networks are trained for each task and dataset separately. The soft-max cross-entropy loss function is used for emotion recognition tasks, while the sigmoid cross-entropy function is employed for AU detection due to the multilabel nature of the latter task.

The accuracy was found to be 54.3% for Single RestNet-50 and 84.2% for SJMT RestNet-50. This method addresses one of the challenges with discrete emotion recognition in the wild, which is the lack of large public labelled datasets.

Exploring Complexity of Facial Dynamics in Autism Spectrum Disorder

Krishnappa Babu et al. (IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 14, NO. 2, APRIL-JUNE 2023)

This work focuses on analyzing the complexity of spontaneous facial dynamics of toddlers with and without ASD. Toddlers watched developmentally-appropriate and engaging movies presented on a smart tablet. Simultaneously, the frontal camera of the tablet was used to record the toddlers' faces, providing the opportunity for the automatic analysis via CV. The facial landmarks' dynamics of the toddlers with ASD versus TD were studied specifically. An iPad-based application (app) was designed that displayed strategically designed, developmentally appropriate short movies involving social and non-social components. The device's front-facing camera was used to record the children's behavior and capture ASD-related features. The children's facial dynamics were exploited from the eyebrows and mouth regions using multiscale entropy (MSE) analysis to study the complexity of such facial landmarks' dynamics. Distinctive landmarks' dynamics were captured in children with ASD, characterized by a significantly higher level of complexity in both the eyebrows and mouth regions when compared to typically-developing children. In Cross-Validation using Decision Tree model the accuracy for each video shown is found to be Video1=77.5%, Video2=74.3%, Video3=73.8%, Video4=67.2%. The study sample has a limited number of ASD participants and did not have sufficient power to determine the impact of demographic characteristics on the results. Other measures of complexity might be more robust than the MSE in their ability to discriminate children with and without ASD.

Real-time facial emotion recognition system among children with autism based on deep learning and IoT

Talaat (F.M. Real-time facial emotion recognition system among children with autism based on deep learning and IoT. Neural Computing & Applications 35, 12717–12728 (2023))

This system proposes an enhanced deep learning (EDL) technique to classify the emotions using convolutional neural network. The proposed emotion detection framework takes the benefit from using fog and IoT to reduce the latency for real-time detection with fast response and to be a location awareness. The architecture outperforms earlier convolutional neural network-based algorithms and does not require any hand-crafted feature extraction. A total of six emotions are detected by the propound system: anger, fear, joy, natural, sadness, and surprise. The training accuracy was found to be 0.963 and validation accuracy 0.88. The limitation of the proposed technique is that it uses small dataset (limited scale) as the large number of real dataset is not available.

Facial Action Unit Detection With Transformers

Jacob et al. (2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR))

The paper outlines an innovative approach for the detection and analysis of facial expressions, with a specific focus on Facial Action Units (FAUs) associated with muscle activations. It highlights the significance of facial expressions as a primary means of conveying nonverbal information, noting that while some expressions are universally understood, others are individualized, necessitating the use of the Facial Action Coding System (FACS). FAU detection is framed as a multi-label binary classification problem, with some approaches considering the degree of FAU activation. The model architecture incorporates attention-based techniques, including separate attention maps for each action unit, and leverages multi-task learning to exploit task relationships. The model's framework involves feature extraction, attention learning, and multi-task modules, with novel loss functions for feature discrimination and multi-label classification. The approach is capable of end-to-end training, achieving state-of-the-art performance on public datasets and undergoing comprehensive evaluation, including ablative studies to assess design choices.

The model shows the best average F1-score of 61.5. The major challenge with the dataset used here is the severity of the class imbalance and the variation in the head pose and expression.

Discriminative Few Shot Learning of Facial Dynamics in Interview Videos for Autism Trait Classification

Zhang et al. (IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 14, NO. 2, APRIL-JUNE 2023)

This paper attempts to fill this gap by developing a novel discriminative few shot learning method to analyze hour-long video data and exploring the fusion of facial dynamics for the trait classification of ASD. The model first extracts the spatio-temporal features of the video and uses the combination of K-SVD with MFA to get more discriminative representations. A few-shot learning module is designed to further improve classification performance It achieves the best performance with an accuracy of 91.72% by fusing the seven selected scenes that are comparable to the standardized diagnostic scales. This Experiment adopts a scene-level feature fusion strategy, which requires manually splitting entire hour-long videos into 15 separate scenes by time markers and extracting facial-dynamics features of each scene.

An Immersive Computer-Mediated Caregiver-Child Interaction System for Young Children with Autism Spectrum Disorder

Nie et al. (IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, VOL. 29, 2021)

The text discusses the development of a computer-mediated system, referred to as the C3I system, designed to enhance the Interactional Joint Attention (IJA) skills of young children with Autism Spectrum Disorder (ASD). The C3I system aims to fill this gap by involving caregivers in the training process. A long video clip is shown to distract the child's attention away from the caregiver. When the child is sufficiently distracted, the caregiver presses a button on a tablet to pause the video. At this point, the expectation is that the child will look back at the caregiver and initiate joint attention. If the child does not respond as expected, the caregiver can press another button to alert the system. In response, the system displays a non-social audio-visual cue, such as a bouncing ball with sound effects, to guide the child's attention either toward the caregiver or one of the target monitors. The system's real-time tracking and response to the child's behavior are central to its operation and its goal of promoting Interactional Joint Attention (IJA) skills in children with ASD.

The feasibility study had only one session per dyad with repetitive trials. In addition, the sample size was small and there was no control group. As such, the results of this feasibility study need to be considered with caution until a larger study verifies its generalizability.

IV. NOVELTY

The project adopts a multimodal approach to emotion recognition in autistic children, combining facial expressions and body gestures for a more comprehensive understanding of their emotions. With the help of GAN (Generative Adversarial Network), the project addresses the common challenge of occluded faces, ensuring that emotion recognition remains same even when parts of the face are concealed. Additionally, the implementation of 3D Convolutional Neural Networks with the attention mapping offers a more refined analysis of emotions, significantly enhancing accuracy. It is designed specifically to cater to the unique challenges of emotion recognition in autistic children. This customization contributes to a profound understanding of how autistic children experience emotions, potentially paving the way for early diagnosis of autism based on emotion recognition. This application of machine learning in the healthcare field holds promise for improving the well-being and support of autistic individuals.

V. ARCHITECTURE

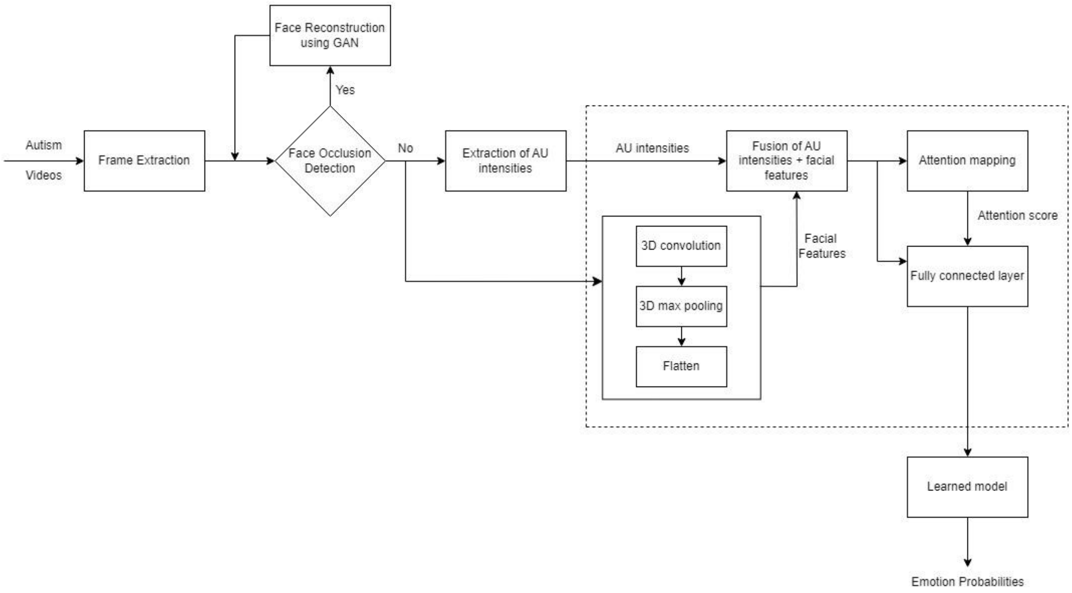


Figure 1: Facial Emotion Detection

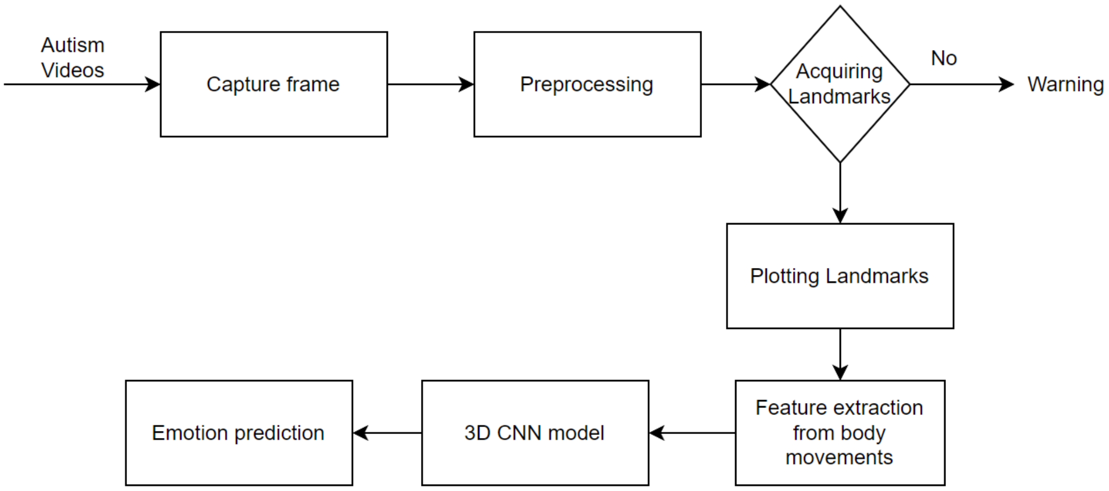


Figure 2: Emotion detection from body languages, hand gestures, etc.

VI. RESULTS

Frame Extraction:

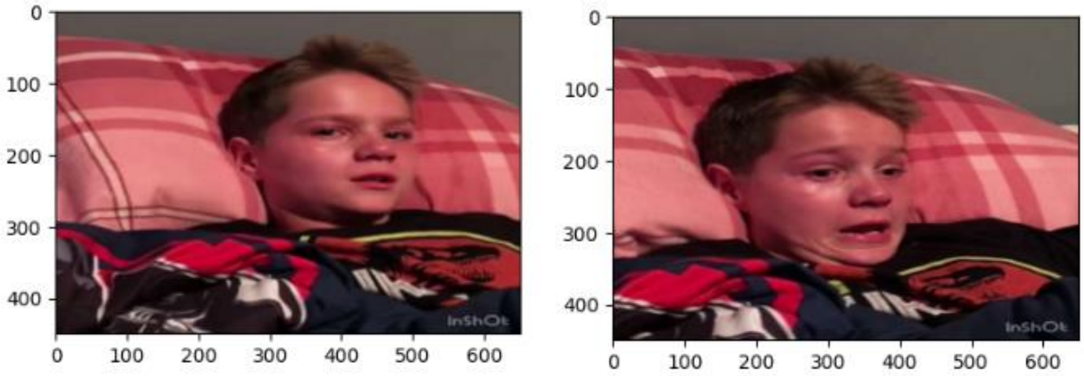


Figure :3 Frames extracted from input video

Preprocessing:

Figure :4 Preprocessed frames extracted from the video

Data Augmentation:

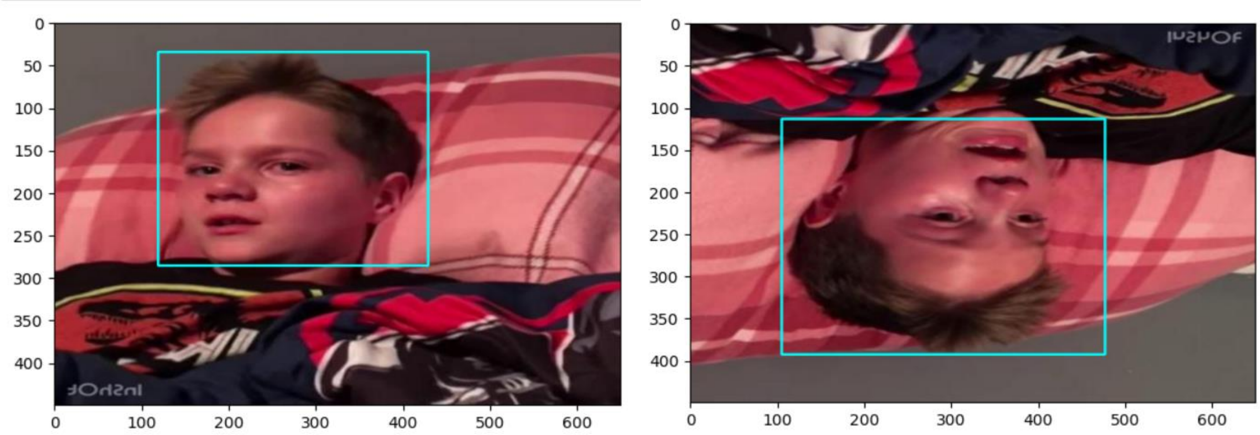


Figure :5 Sample augmented frames

Face Recognition Output:

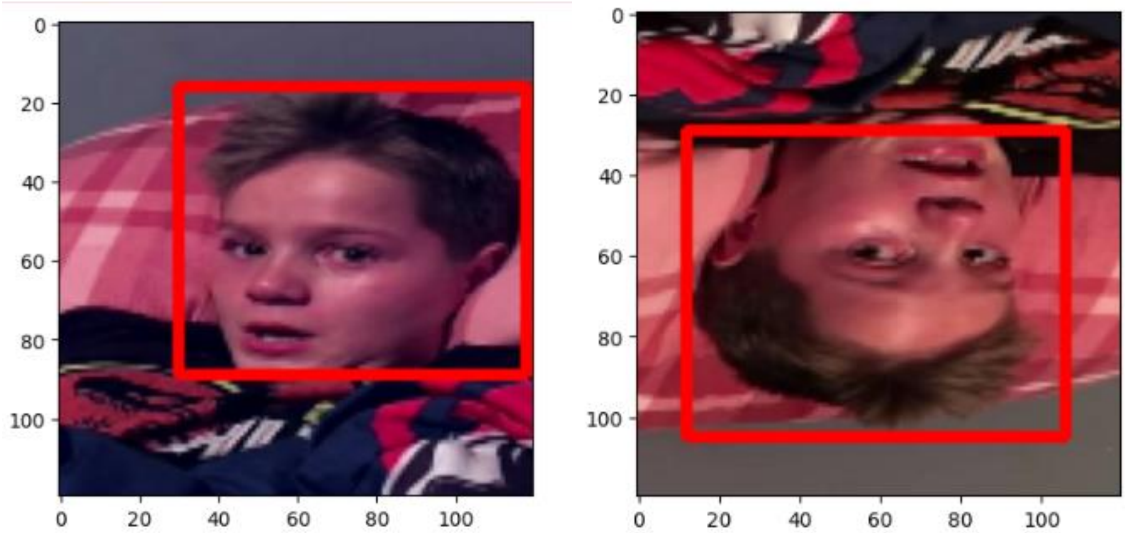
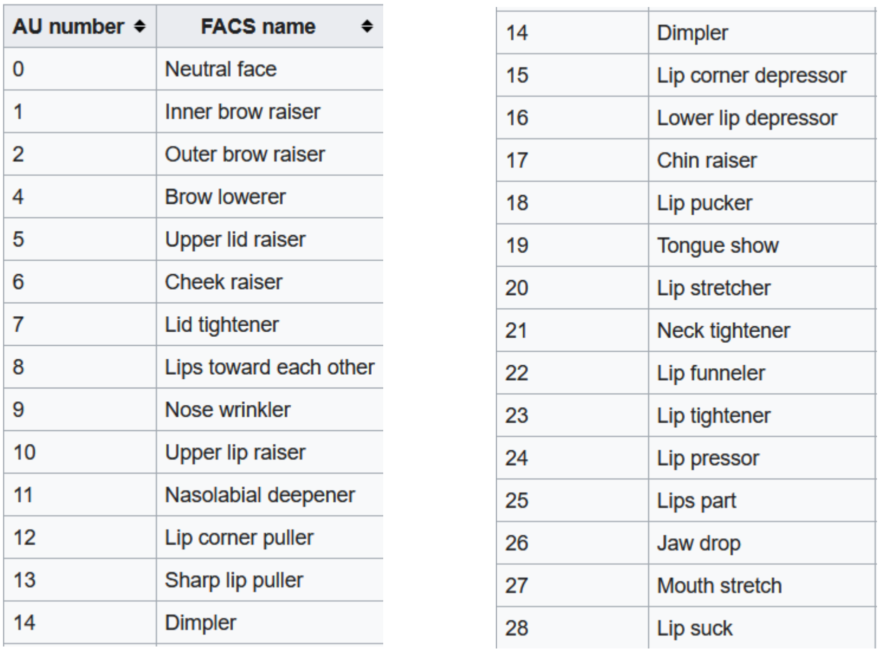


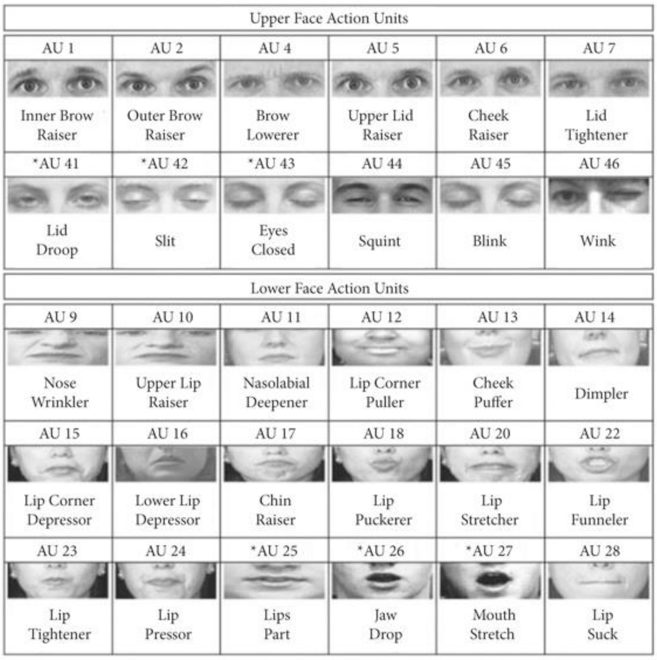
Figure :6 Output of the face recognition model

List of AUs (with underlying facial muscles):



AU number ⬍	FACS name ⬍
0	Neutral face
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowerer
5	Upper lid raiser
6	Cheek raiser
7	Lid tightener
8	Lips toward each other
9	Nose wrinkler
10	Upper lip raiser
11	Nasolabial deepener
12	Lip corner puller
13	Sharp lip puller
14	Dimpler
14	Dimpler
15	Lip corner depressor
16	Lower lip depressor
17	Chin raiser
18	Lip pucker
19	Tongue show
20	Lip stretcher
21	Neck tightener
22	Lip funneler
23	Lip tightener
24	Lip pressor
25	Lips part
26	Jaw drop
27	Mouth stretch
28	Lip suck

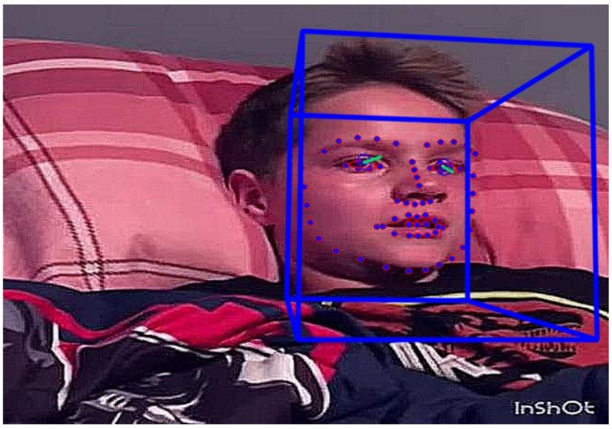
Figure :7 List of Action Units with respect to facial muscles



Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

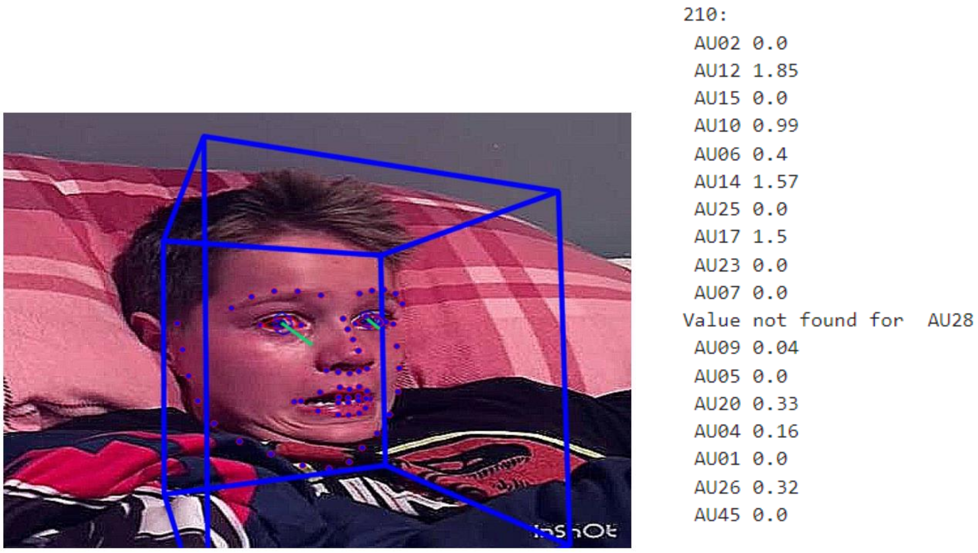
Figure :8 Pictorial representation of action units

Facial Action Unit detection:



```
1:
 AU02 0.0
 AU12 0.1
 AU15 0.1
 AU10 0.0
 AU06 0.0
 AU14 0.14
 AU25 1.33
 AU17 0.18
 AU23 0.0
 AU07 0.0
Value not found for  AU28
 AU09 0.0
 AU05 0.16
 AU20 0.0
 AU04 0.03
 AU01 0.23
 AU26 0.0
 AU45 0.0
```

Figure :9 Output of facial action unit detection



```
210:
 AU02 0.0
 AU12 1.85
 AU15 0.0
 AU10 0.99
 AU06 0.4
 AU14 1.57
 AU25 0.0
 AU17 1.5
 AU23 0.0
 AU07 0.0
Value not found for  AU28
 AU09 0.04
 AU05 0.0
 AU20 0.33
 AU04 0.16
 AU01 0.0
 AU26 0.32
 AU45 0.0
```

Figure :10 Output of facial action unit detection

VII. EVALUATION

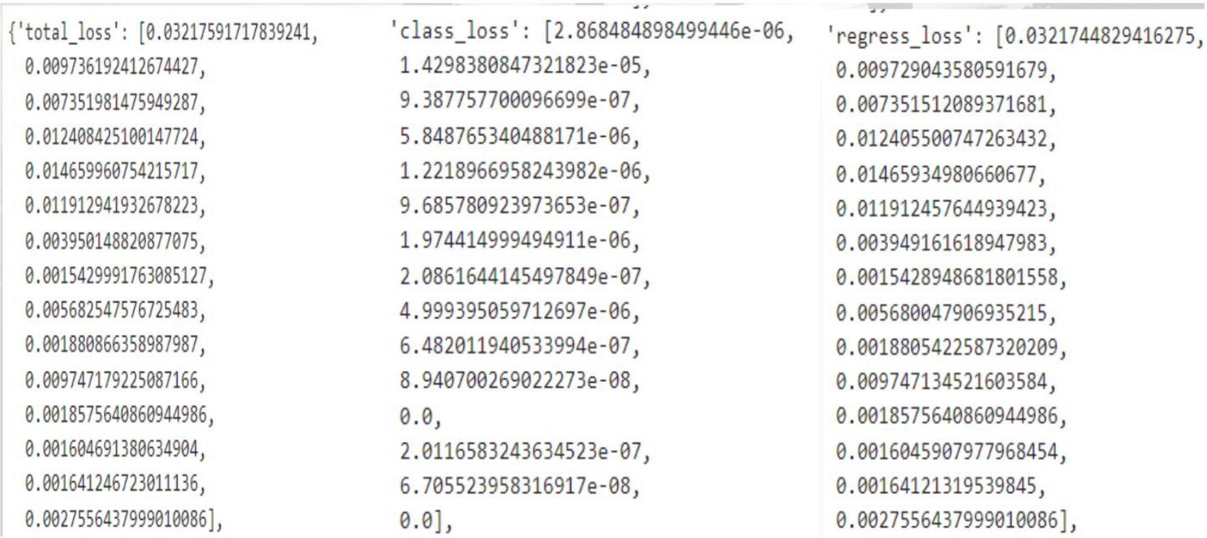
Loss Equation:

$$\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right) \right.$$

Figure :9 Loss equations used for regression loss

Losses in Face recognition:



```
{'total_loss': [0.03217591717839241,      'class_loss': [2.868484898499446e-06,      'regress_loss': [0.0321744829416275,
  0.009736192412674427,                     1.4298380847321823e-05,                     0.009729043580591679,
  0.007351981475949287,                     9.387757700096699e-07,                      0.007351512089371681,
  0.012408425100147724,                     5.848765340488171e-06,                      0.012405500747263432,
  0.014659960754215717,                     1.2218966958243982e-06,                     0.01465934980660677,
  0.011912941932678223,                     9.685780923973653e-07,                      0.011912457644939423,
  0.003950148820877075,                     1.974414999494911e-06,                      0.003949161618947983,
  0.0015429991763085127,                    2.0861644145497849e-07,                     0.001542894868181558,
  0.005682547576725483,                     4.999395059712697e-06,                      0.005680047906935215,
  0.001880866358987987,                     6.482011940533994e-07,                      0.0018805422587320209,
  0.009747179225087166,                     8.940700269022273e-08,                      0.009747134521603584,
  0.0018575640860944986,                    0.0,                                        0.0018575640860944986,
  0.001604691380634904,                     2.0116583243634523e-07,                     0.0016045907977968454,
  0.001641246723011136,                     6.705523958316917e-08,                      0.00164121319539845,
  0.002755643799010086],                    0.0],                                       0.002755643799010086],
```

Figure :10 Losses in the face recognition model

Graphical representation of losses in Face recognition

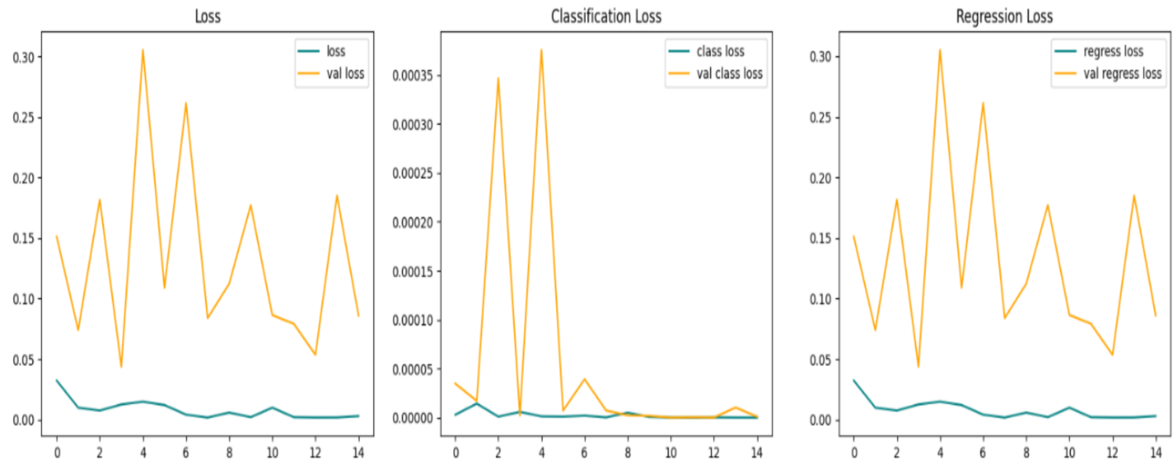


Figure :11 Graphical representation of the losses in the face recognition

VIII. CONCLUSION

In conclusion, this project aims to enhance the emotional well-being of autistic children by creating a specialized system for emotion detection. The use of GANs helps to overcome challenges like occluded faces to ensure that emotions can be recognized even when parts of the face are hidden. This system combines facial expression analysis, body movement tracking, and landmark detection to provide a comprehensive view of the child's emotional state. This not only aids in better understanding how autistic children experience emotions but also equips educators, therapists, and caregivers with valuable insights to provide effective emotional support.

IX. REFERENCES:

[1] Z. Shao, Y. Zhou, J. Cai, H. Zhu and R. Yao, "Facial Action Unit Detection via Adaptive Attention and Relation," in IEEE Transactions on Image Processing, vol. 32, pp. 3354-3366, 2023, doi: 10.1109/TIP.2023.3277794. Link

[2] G. Pons and D. Masip, "Multitask, Multilabel, and Multidomain Learning With Convolutional Networks for Emotion Recognition," in IEEE Transactions on Cybernetics, vol. 52, no. 6, pp. 4764-4771, June 2022, doi: 10.1109/TCYB.2020.3036935. Link

[3] Yao, L., Wan, Y., Ni, H. et al. Action unit classification for facial expression recognition using active learning and SVM. Multimed Tools Appl 80, 24287–24301 (2021). Link

[4] M. K. Tellamekala, Ö. Sümer, B. W. Schuller, E. André, T. Giesbrecht and M. Valstar, "Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2023.3280530. Link

[5] P. R. K. Babu et al., "Exploring Complexity of Facial Dynamics in Autism Spectrum Disorder," in IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 919-930, 1 April-June 2023, doi: 10.1109/TAFFC.2021.3113876. Link

[6] M. Talaat, Fatma. (2023). Real-time facial emotion recognition system among children with autism based on deep learning and IoT. Neural Computing and Applications. 35. 10.1007/s00521-023-08372-9. Link

[7] Geethu Miriam Jacob, Bjorn Stenger; Facial Action Unit Detection With Transformers, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7680-7689 Link

[8] N. Zhang, M. Ruan, S. Wang, L. Paul and X. Li, "Discriminative Few Shot Learning of Facial Dynamics in Interview Videos for Autism Trait Classification," in IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1110-1124, 1 April-June 2023, doi: 10.1109/TAFFC.2022.3178946. Link

[9] Rajaram, Santhoshkumar & Geetha, M.. (2019). Deep Learning Approach for Emotion Recognition from Human Body Movements with Feedforward Deep Convolution Neural Networks. Procedia Computer Science. 152. 158-165. 10.1016/j.procs.2019.05.038. Link

[10] Zhuang W, Chen L, Hong C, Liang Y, Wu K. FT-GAN: Face Transformation with Key Points Alignment for Pose-Invariant Face Recognition. *Electronics*. 2019; 8(7):807. https://doi.org/10.3390/electronics8070807