REWARD DRIVEN EMOTION DETECTION IN AUTISM SPECTRUM DISORDER WITH ATTENTION MECHANISM

IT5712 Project-1 Report

Submitted by

Rahul Prasanth D	2020506070
Bala Natesh R M	2020506018
Sanjay G	2020506080

Under the supervision of

Dr. J. Dhalia Sweetlin

In partial fulfillment for the award of the degree
of
BACHELOR OF TECHNOLOGY
in
INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

MADRAS INSTITUTE OF TECHNOLOGY CAMPUS

ANNA UNIVERSITY, CHENNAI – 600044

NOVEMBER 2023

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled **“REWARD DRIVEN EMOTION DETECTION IN AUTISM SPECTRUM DISORDER WITH ATTENTION MECHANISM”** is the bonafide work of Rahul Prasanth D (2020506070), Bala Natesh R M (2020506018) and Sanjay G (2020506080) who carried out the project work under my supervision.

Signature

Dr. M. R. Sumalatha

HEAD OF THE DEPARTMENT

Professor
Department of Information Technology
MIT Campus, Anna University
Chennai – 600044

Signature

Dr. J. Dhalia Sweetlin

SUPERVISOR

Associate Professor
Department of Information Technology
MIT Campus, Anna University
Chennai – 600044

ACKNOWLEDGEMENT

It is essential to mention the names of the people, whose guidance and encouragement made us accomplish this project.

We thank our project supervisor **Dr. J. Dhalia Sweetlin**, Associate Professor of the Department, Department of Information Technology, MIT Campus, for providing invaluable support and assistance with encouragement which aided to complete this project.

We are very thankful to the panel members **Dr. P. AnandhaKumar** and **Dr. M. Hemalatha,** Department of Information Technology, MIT Campus for their invaluable feedback in reviews.

Our sincere thanks to **Dr. M. R. Sumalatha**, Head of the Department of Information Technology, MIT Campus for catering all our needs giving out limitless support throughout the project phase.

We express our gratitude and sincere thanks to our respected Dean of MIT Campus, **Dr. J. Prakash**, for providing excellent computing facilities throughout the project.

RAHUL PRASANTH D	**2020506070**
BALA NATESH R M	**2020506018**
SANJAY G	**2020506080**

ABSTRACT

Emotion recognition in ASD individuals is particularly challenging due to variations in expressive behaviours. This project aims to address these challenges by developing a specialized system for comprehensive emotion recognition in autistic children. The proposed system utilizes a multimodal approach, incorporating facial expressions and their body movements to provide an elaborate view of emotional states. It takes the videos of the autistic children and the extracted frames after preprocessing is used to extract the facial landmark points along with the body landmark points. The extracted landmark points are given as an input to the LSTM model for extraction of temporal features. Then the attention mechanism is implemented in the features extracted from the LSTM model. The attention scores of the features are given to the fully connected layer along with the reward functions to enhance the performance of the prediction model. Based on the probabilities of each emotion from the fully connected layer, the emotion is classified. The system aims to provide valuable insights into the emotional experiences of autistic children, aiding educators, therapists, and caregivers in offering tailored support and interventions. The proposed work improves communication and emotional well-being in autistic children.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASD	Autism Spectrum Disorder
GAN	Generative Adversarial Network
LSTM	Long Short Term Memory

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Understanding and expressing emotions can be particularly challenging for children with Autism Spectrum Disorder (ASD). Traditional methods of recognizing emotions often miss the complex cues that are essential for effective communication in this population. This project addresses this gap by introducing a specialized system designed to decode the emotions of autistic children in a comprehensive way. This work goes beyond just facial expressions, encompassing body movements to provide a elaborate view of emotional states. Autistic children often communicate emotions differently, and by considering these diverse modalities, this system aims to create a more accurate and sensitive understanding. The proposed work acknowledges the everyday scenarios where faces may be partially hidden, it also be taken into account. The system is designed to be a bridge, helping educators, therapists, and caregivers connect more effectively with these children. By tapping into where and when emotions happen on their faces, coupled with specific facial expressions and body language, tools can be created that offers practical and meaningful support.

1.2 RESEARCH CHALLENGES

Krishnappa et al., 2023 suggested that developing the emotion recognition system in autistic children have intricate challenges that demand innovative solutions. The main challenge is the issue of facial occlusion, where traditional methods often struggle to precisely capture facial expressions when faces are partially hidden. Another critical challenge lies in the diverse and complex nature of emotional expression among autistic children. Conventional emotion recognition models, trained on datasets predominantly composed of neurotypical individuals, may not adequately capture the unique ways in which autistic children convey emotions.

Autistic individuals often exhibit a broad spectrum of expressive behaviours, and the lack of diversity in training data can lead to models that struggle to generalize across this spectrum. Moreover, the scarcity of labelled data specific to autistic children poses a significant hurdle. The complex nature of emotions requires a substantial amount of diverse and accurately labelled data for effective model training. Acquiring such data is challenging due to privacy concerns, ethical considerations, and the need for expertise in labelling the complex emotional expressions.

1.3 OBJECTIVE

The primary objective of this system is to develop a specialized emotion recognition system tuned to the complex expressive behaviours of autistic children. By exploring the facial expressions and body movements, the goal is to construct a comprehensive tool that works beyond the traditional recognition methods. In addition to this, the system aims to address the challenge of occluded faces for the emotion prediction. It aspires to construct a bridge connecting educators, therapists, and caregivers to the emotional experiences of autistic children

1.4 SCOPE OF THE PROJECT

The scope of this project encompasses the development of a specialized emotion recognition system tailored to the complex expressive behaviours of autistic children. The project will focus on exploring facial expressions and body movements as key modalities for emotion recognition. A significant aspect of the project's scope involves addressing the challenge of occluded faces by incorporating methods for face reconstruction when parts of the face are hidden. However, it's important to note that the project concentrates specifically on enhancing emotion recognition to improve communication and support.

1.5 CONTRIBUTION

The project contributes by developing a specialized emotion recognition system for autistic children, incorporating a reward mechanism to enhance performance of the model. Addressing occluded faces and focusing on facial expressions and body movements, the system provides valuable insights to educators, therapists, and caregivers, ultimately improving understanding and support for the emotional experiences of autistic children.

1.6 ATTENTION MECHANISM

The attention mechanism integrated into this system plays an important role in the process of emotion recognition with the help of the facial expressions and the body movements. As the autistic children express diverse and unique emotions, the attention mechanism helps to find the hidden features or the diverse features by assigning the high attention scores. The attention mechanism operates like a dynamic filter, which allows the system to selectively focus on specific regions of interest within facial expressions and body language. This system achieves an elevated level of sensitivity and precision, ensuring a more elaborate understanding of the emotions. Moreover, the attention mechanism proves to be a strategic solution to the challenges posed by hidden features, compensating for potential limitations in data input like hidden faces, etc. (Jacob et al., 2023) underscores the significance of incorporating attention mechanism in emotion recognition systems tailored for autism. The attention mechanism, as evidenced by the research, significantly enhances the overall effectiveness of the system.

1.7 REWARD FUNCTION

The proposed work introduces a reward function during which acts as a guide for the model. Unlike traditional approaches, the reward function plays a transformative role by encouraging the model's ability to correctly identify and interpret emotions, helps the model to prioritize precision. By positively reinforcing correct identifications, the system not only refines its understanding of the diverse emotional expressions inherent to ASD but also adapts dynamically to complex cues. This reward mechanism essentially guides the model towards more reliable and complex emotion detection, contributing to an enhanced and refined system tailored to the complexities of Autism Spectrum Disorder.

1.8 ORGANIZATION OF THESIS

The rest of the thesis is organized as follows. Chapter 2 presents the literature survey on emotion detection methodologies. Chapter 3 models the architecture and system design, outlining the architecture and design of the proposed system. Chapter 4 explains about the implementation details, explaining the specifications and environment. The results achieved are presented in Chapter 5. Chapter 6 presents the conclusion and some possible avenues for future research on the topic.

CHAPTER 2
LITERATURE SURVEY

A Survey conducted on the emotions of the autistic children and their emotion dynamics. The dynamic nature of their emotions, problems in determining the emotions, techniques to determine the emotions are given below.

2.1 FACIAL DYNAMICS OF AUTISTIC CHILDREN

Toddlers were shown developmentally-appropriate and engaging movies presented on a smart tablet whose, frontal camera was used to record their faces, providing the opportunity for the automatic analysis via CV (Krishnappa et al., 2023). The facial landmarks' dynamics of the toddlers were studied specifically. 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.

2.2 FACIAL EMOTION DETECTION:

The proposed Face Detection Convolutional Neural Network (FDCNN) model exhibits a structured architecture tailored for emotion recognition in images derived from video frames (Santoshkumar et al., 2019). With an input size of 150x150x3, representing the RGB channels, the model incorporates three convolutional layers, each applying 3x3 filters. The initial layer employs 32 filters, leading to 32 stacked feature maps, followed by max pooling to reduce spatial dimensions. Subsequent layers involve 64 and 128 filters, with corresponding max pooling operations, resulting in feature maps of dimensions (75x75x32), (37x37x64), and (18x18x128), respectively. The model culminates in fully connected layers, featuring 512 hidden units and an output layer with 15 neurons, aligning with the number of emotion classes. Training and validation sets are generated by converting input videos into frames, facilitating the model's training and evaluation. This architecture conforms to the standard convolutional neural network paradigm for image classification tasks, utilizing convolutional layers to capture hierarchical features and max pooling for spatial downsampling. The fully connected layers contribute to the extraction of high-level representations before producing emotion predictions. The model's design, combining convolutional and fully connected layers, underscores its effectiveness in discerning complex patterns within facial expressions, making it well-suited for applications in emotion recognition from video frames.

Another approach called Discriminative Few Shot Learning which is proposed by Zhang et al., 2023. The FSL system, when combined with the fusion of feature levels from each scene, achieves an impressive accuracy of 91.72% on the Caltech ADOS video data. The scene-level fusion, reveals insights into the unequal distribution of diagnostic information across different scenes and asserting that ASD is a complex condition requiring nuanced phenotyping beyond conventional classification categories. The model begins by extracting spatio-temporal features from the video, employing a combination of K-SVD with Marginal Fisher Analysis (MFA) to derive more discriminative representations. The scene-level feature fusion strategy, requires manually splitting entire hour-long videos into 15 separate scenes by time markers and extracting facial-dynamics features of each scene.

2.3 FACIAL OCCLUSION:

Most of the facial emotion detection models face the issue of occlusion in face which affects the models performance in predicting the emotions. So it can be solved by facial reconstruction and complete face recovery techniques. The Deep Cascade

Guidance Learning method involves a three-stage guidance learning scheme-occlusion detection, face parsing, and face reconstruction. The first two stages are trained on both synthesized and real data domains, enabling domain-agnostic guidance for the subsequent reconstruction stage and effectively mitigating the prevalent domain gap issue. By disentangling input information into domain-agnostic and appearance inputs, the cascade guidance learning model significantly reduces reliance on domain-sensitive appearance details, resulting in a substantial enhancement in the performance of face reconstruction on real-world images. Two more reference modules based on masked attention models are used that demonstrate both effectiveness and efficiency in inpainting occluded facial parts. This work's performance is compared with the standard model like RCPR, HRNet on different standard datasets like COFW and 300W, the proposed model gives the better reconstructed image with low normalized mean error between the occluded face and the reconstructed face (Ni Zhang et al., 2023)

2.4 ATTENTION MECHANISM

The adaptive attention regression network, integrated with local attention predefinition and global attention learning, captures both predefined dependencies by landmarks in strongly correlated regions (regions which going to make a greater impact) and facial globally distributed dependencies in weakly correlated regions (Shao et al., 2023). An adaptive spatio-temporal graph convolutional network simultaneously reasons the specific pattern of each AU, the inter-dependencies among AUs, as well as the temporal correlations. 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.

2.5 SUMMARY OF THE LITERATURE SURVEY

The survey explores a diverse array of methodologies for facial emotion recognition in individuals with Autism Spectrum Disorder (ASD). The studies highlight the importance of complex phenotyping, acknowledging the complexity of ASD and emphasizing the need for adaptive and comprehensive models. Various models, such as the Face Detection Convolutional Neural Network (FDCNN), Few-Shot Learning (FSL) systems, and discriminative Facial Action Unit (FAU) detection, showcase promising results in capturing and understanding facial expressions in normal individuals. The integration of attention mechanisms, multi-task learning, and innovative loss functions underscores the continuous effort to improve model performance and address challenges like class imbalance and variations in datasets. Moreover, cascade guidance learning, and adaptive attention regression networks aim to enhance robustness in real-world scenarios, addressing issues related to occlusions and misalignment errors. Challenges include limited sample sizes, domain gaps, misalignment errors, insufficient labelled datasets, variability in emotion expression, ethical considerations, and the need for real-time processing

CHAPTER 3

SYSTEM ARCHITECTURE AND DESIGN

3.1 SYSTEM ARCHITECTURE

The proposed project architecture is designed as a robust pipeline for the prediction of emotions of Autism Spectrum Disorder (ASD) children. The process initiates with Autism Videos serving as the primary dataset, and subsequent Frame Extraction facilitates the breakdown of video content into individual frames to enable granular analysis. Preprocessing steps ensure the frames are suitably prepared for subsequent analysis, including resizing, adjusting the sharpness level and the brightness in the frames. The Facial Landmark Extraction captures 468 crucial facial features, while Body Landmark Extraction concurrently gathers key body landmark points which includes the 21 landmark points in each hand, legs and so on. The total number of body pose landmarks are 33 which is captured in the body landmark extraction. Then the extracted landmark points are concatenated and used for the further predictions. The architecture further integrates an LSTM (Long Short-Term Memory) layer, a recurrent neural network designed to comprehend temporal dependencies and patterns in the sequence of frames, enhancing the system's ability to understand dynamic facial and body expressions over time. Attention Mapping and the Reward Function are seamlessly incorporated into the LSTM layer, providing dual functionality to enhance focus on critical features and give a positive reinforcement during training. The Fully Connected Layer is connected to the Attention Layer, suggesting that the fully connected layer is incorporating information that has been selectively attended to by the attention mechanism. The reward is given in the form of the loss and the reward function consider the parameters like attention scores, predicted values, normalization approaches, regularization techniques, etc. The reward during loss implies that, during training, the model is rewarded for making correct predictions or recognizing important patterns, enhancing its ability to learn and improve over successive iterations. In the final layer of the network architecture, the model outputs probabilities for six distinct emotion classes: anger, happy, surprise, neutral, fear, and sad. This output layer represents a comprehensive classification scenario for emotion recognition, assigning a probability distribution across these six classes for each input. Practical interpretation involves considering the class with the highest probability as the predicted emotion for a given input. This probabilistic approach provides a better understanding of the model's confidence in its predictions, allowing for a more detailed assessment of uncertainty or ambiguity in emotion recognition. The proposed work is shown in the Figure 3.1

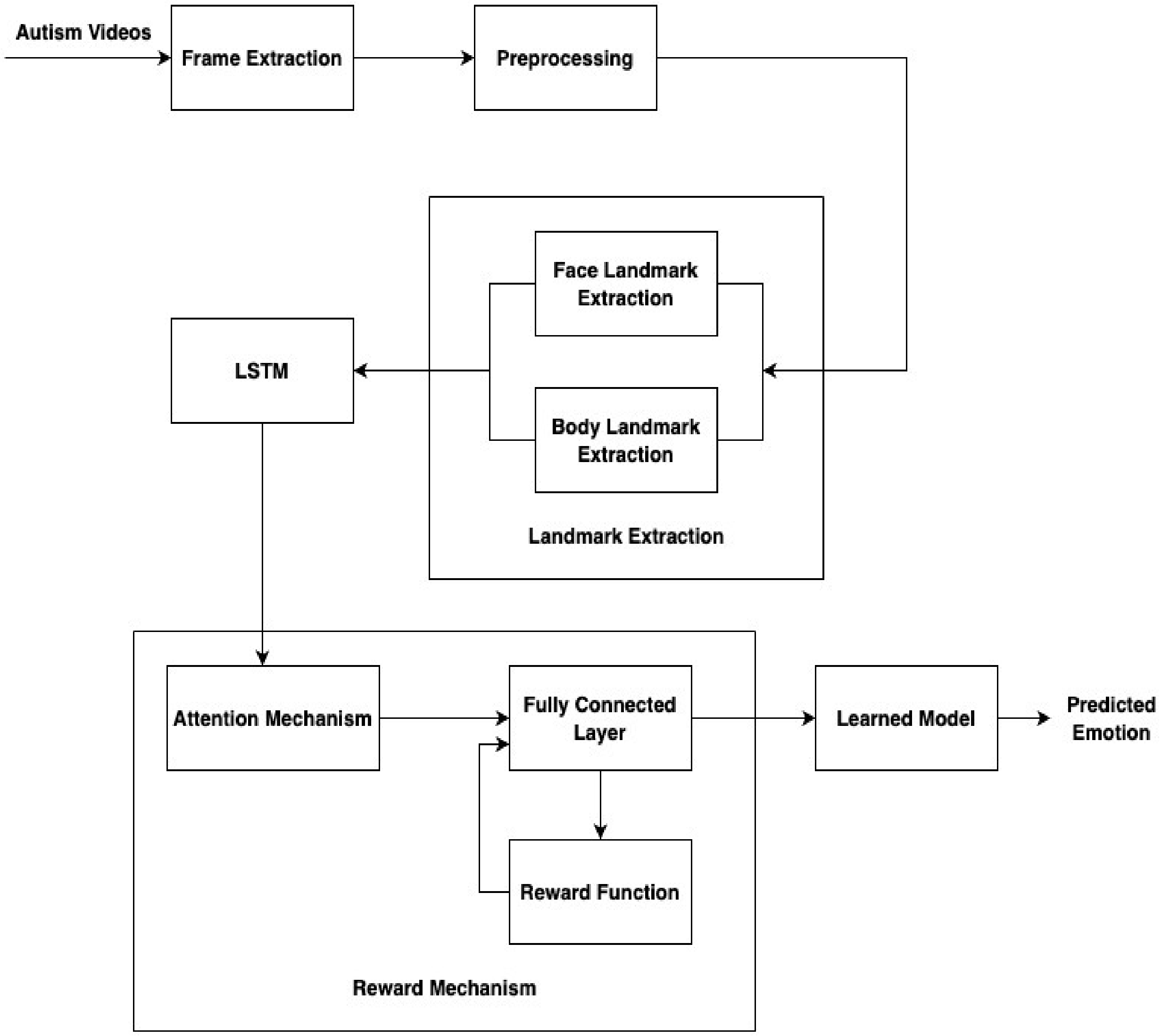


Figure 3.1 – System Architecture

3.2 LSTM

Long Short-Term Memory (LSTM) networks, a subtype of recurrent neural networks (RNNs), play a major role in emotion detection for the individuals with Autism Spectrum Disorder (ASD). LSTMs are good at capturing the temporal dynamics inherent in facial expressions along with their body movements, making them an efficient one for analysing evolving emotional states over time. The distinctive feature of LSTMs lies in their ability to overcome the vanishing gradient problem, ensuring effective learning and retention of long-term dependencies in sequential data. Within the framework of emotion detection, LSTMs process sequential input data, enabling them to consider the temporal evolution of facial expressions. The memory cells inherent in LSTMs facilitate the retention of information over extended periods, providing the model with the capacity to discern nuanced changes in emotional states. The incorporation of gating mechanisms, including input, output, and forget gates, allows LSTMs to selectively retain relevant information, and helps to transfer the content to the next hidden state, enhancing their capacity to understand the complexities of evolving emotions. In the integration with emotion detection, LSTMs contribute a crucial temporal context. They consider not only the current frame but also the sequence of expressions leading up to it, offering a more comprehensive understanding of emotional dynamics. The benefits of LSTMs in emotion detection lie in their temporal sensitivity, capturing subtle changes in facial expressions, and their ability to understand long-term dependencies, providing insights into the progression of emotions. Integrating attention mechanism along with the reward function enhances the overall accuracy and depth of emotion detection in individuals with ASD**.**

3.2.1 ATTENTION MECHANISM

The attention mechanism is in emotion detection for individuals with Autism Spectrum Disorder (ASD) which enables selective focus on crucial facial regions, emphasizing expressive features and promoting contextual understanding. Its dynamic adaptation allocates attention weights based on evolving emotional contexts on the temporal data. Integrated with facial landmarks analysis and body landmark analysis, it prioritizes regions essential for expressing emotions. When combined with Long Short-Term Memory (LSTM) networks, attention mechanisms contribute context-aware temporal analysis of facial expressions along with the hand and body movements.

3.2.2 REWARD FUNCTION

The incorporation of a reward function in the emotion detection system transforms the training process by providing dynamic feedback during each iteration based on the model's predictions. Unlike conventional methods, this approach adapts in real-time, offering rewards contingent on precise emotion recognition. Tailored for the complexity of emotions, the reward function ensures individualized and adaptive learning, allowing the model to refine its predictions iteratively. During training, this mechanism evaluates the model's predictions, rewards for accurate emotion recognition and guiding the model towards improved performance. Its iterative nature enables continual refinement, facilitating the model's adaptation to diverse emotional expressions over successive training cycles. Integrated into the training process, the reward function augments traditional methodologies, optimizing the model's parameters and promoting accurate emotion recognition. The proposed work uses the four different reward function, based on the performance of each reward function, the best performing function will be used for the prediction of the emotion. The real-time feedback enhances the model's ability to capture complex emotional states, making it an important component in the adaptive landscape of emotion detection.

CHAPTER 4

ALGORITHM DEVELOPMENT AND IMPLEMENTATION

4.1 LSTM IMPLEMENTATION

The proposed work's network architecture comprises multiple layers designed for sequence processing and attention-based feature extraction. The model begins with an LSTM layer, configured for 30 time steps and 1662 features in each sequence which includes the facial landmark points, body pose landmarks, hand landmark points, followed by a attention layer to emphasize relevant features. Two additional LSTM layers capture temporal dependencies, leading to a final Dense layer stack for classification of six different emotions. The architecture is structured to leverage the strengths of LSTM units in sequence learning, with attention mechanisms enhancing the network's focus on crucial information and the reward function to penalize the incorrect predictions. The model aims to discern patterns within input sequences, crucial for tasks like emotion detection, where temporal features are vital. Optimization is achieved through the use of rectified linear unit (ReLU) activations, and the final softmax layer facilitates multi-class classification. This comprehensive design seeks to extract meaningful representations from sequential data, leveraging attention mechanisms for enhanced discrimination, making it particularly suitable for tasks requiring understanding of temporal dynamics, such as emotion recognition in diverse datasets

4.2 ATTENTION MECHANISM

The attention mechanism algorithm involves several key steps, and a commonly used version is the scaled dot-product attention. The scaled dot-product attention is used because they are computationally efficient, particularly when implemented with matrix operations. The attention scores generated by this mechanism can provide insights into which parts of the input sequence are more relevant for a given output.

The scaling factor helps to stabilize the gradients during training. This is particularly important in deep neural networks, where vanishing gradients can be a challenge.

$$Attention(q,k,v)=Softmax(\frac{q.(k^T)}{\sqrt{d}}).\ v \qquad (4.1)$$

Where the mathematical representation of the scalar dot product attention mechanism is given by the equation 4.1 which consists of the terms q, k, v. The term ‘q’ refers to the query vector which represents the information the model seeks to retrieve from the input sequence. The term ‘k’ refers to the vectors which contains information about the elements in the sequence and is used to compute the similarity with the query. Then the term ‘v’ holds the information associated with each element in the sequence and the term ‘d’ represents the key vector dimension which as a scaling factor. The softmax function normalizes the similarity scores, producing attention weights that indicate the importance of each element in the sequence. The scaling factor prevents the dot products from becoming too large, contributing to more stable and efficient training. The dot product measures the similarity between the query and key vectors, allowing the model to capture dependencies and relationships across the entire input sequence.

4.3 REWARD FUNCTION

In this proposed work, four distinct reward functions are integrated into the training process. These reward mechanisms are strategically employed during loss propagation, contributing to the model's learning process. By incorporating diverse reward signals, it aims to reinforce precise emotion recognition and enhance the overall performance of the system. The reward function consists of concepts like normalization, regularization, non linearity, overfitting problems, etc. If the reward value given by the reward function is large, it makes the propagation of smaller loss value. If the predictions are incorrect then larger reward values are propagated which acts like a penalty for the incorrect predictions.

4.3.1 FUNCTION 1:

The rewards are calculated using the equation 4.2 and the rewards are given in form of losses that is the positive reward makes the loss value lesser and the negative reward makes the loss value greater than the existing loss value.

$$Reward = (pred\_correct) - (0.1 * diversity\_penalty) - (0.2 * weighted\_penalty\ ) + 0.3 \tag{4.2}$$

$$pred\_correct = \sum_{i=1}^{n} y_{true,i} - y_{pred,i} \tag{4.3}$$

$$diversity\_penalty = ||mean\_pred\_axis\_0 - mean\_pred\_axis\_1|| \tag{4.4}$$

$$weighted\_penalty = \sum_{i=1}^{n} (y_{true,i} - y_{pred,i})^2 * weight[i] \tag{4.5}$$

Where $pred\_correct$ in the equation (4.2) is the loss between the true values and the predicted values which is computed using the mathematical equation (4.3). The term y_{true} represents the truth values and y_{pred} represents the predicted values. Then $diversity\_penalty$ is to increase the diversity in the prediction of emotions to prevent the overfitting problem and it is computed by the equation (4.4). It considers the mean predictions along the different axes that is one axis gives the temporal predictions of the emotions which is represented by $mean\_pred\_axis\_0$ and the other axis consists of the probabilities of each emotions which is represented by $mean\_pred\_axis\_1$. By considering the difference between the mean predictions in both axes, the diversity is included in the training. It prevents certain patterns from dominating the model's output along specific axes. It helps to promote a more balanced and diverse set of predictions. The $weighted\_penalty$ accounts for the mistakes made by the model, the difference between the predicted emotion and true emotion represents the errors for each class. The penalties are weighted by the corresponding class weights which is given by the equation (4.5). The weights for each class is given the weight vector which is predefined. All these parameters are combined along with the bonus value which is used to normalize the reward value, the reward will be calculated.

4.3.2 FUNCTION 2:

The Reward function 2 is the modified version of the equation (4.2), it is also used to update the loss values for increasing the efficiency of the model by considering the mean value of the attention score. Equation (4.6) gives the mathematical representation.

$$Reward = (pred\_correct) - (0.1 * diversity\_penalty) - (0.3 * weighted\_penalty) + 0.4 + (0.2 * mean(attention\_score)) \tag{4.6}$$

Where the $pred\_correct$ represents the correctly predicted value which is already given in the equation (4.3). The $diversity\_penalty$ is given in the equation (4.4) and the $weighted\_penalty$ also same as the reward function 1, its mathematical representation is given by the equation (4.5). The difference between the reward function 1 and reward function 2 is the consideration of the average of attention score. The $attention\_score$ is computed using the equation (4.1). With the help of attention score's average, the model can infer how the attention score is distributed over the data.

4.3.3 FUNCTION 3:

The reward function 3 is the modified version of equation (4.6) where additional parameters are included to increase the performance of the model. The modified reward function in given in the equation (4.7) which consists of attention score for the $weighted\_penalty$ and a new term is included to balance the reward.

$$Reward = (pred\_correct) - (0.1 * diversity\_penalty) + (0.2 * average\_attention\_score\ ) - (0.3 * weight\_penalty\ ) + n\_term \tag{4.7}$$

$$pred\_correct = \sum_{i=1}^{n} y_{true,i} * y_{pred,i} * (1 + \log(external)) \tag{4.8}$$

$$weight\_penalty = \sum_{i=1}^{n} \left(y_{true,i} - y_{pred,i}\right)^2 * weight[i] * attention\_score[i] \tag{4.9}$$

$$n\_term = (0.2 * external) + (0.3 * mean(\sqrt{attention\_score}) \tag{4.10}$$

Where the $pred\_correct$ is the cross entropy between the true value y_{true} and the predicted value y_{pred} along with a logarithmic transformation by considering the $external$ metric which is given in the equation (4.8). Then $diversity\_penalty$ encourages the model to make the diverse predictions by penalizing the similarity between the mean prediction along different axes and the mathematical representation is given by the equation (4.4). The $weight\_penalty$ is to penalize predictions errors in a weighted manner which takes both class weights given by predefined weight vector and the attention score computed using equation (4.1). The $weight\_penalty$ is mathematically represented in equation (4.9). Then the n_term acts as a bonus to balance the reward with the help of the $external$ metric and the attention score and it is given by the mathematical equation (4.10).

4.3.4 FUNCTION 4

The reward function 4 includes the temporal regularization components to focus on the temporal dynamic by considering the predicted values and adding the parameters to include the non linearity. This is mathematically represented in equation (4.11).

$$Reward = (pred\_correct) - (0.1 * diversity\_penalty) - (0.3 * weighted\_penalty) + (0.2 * avg\_attention) + n\_term + temporal\_reg \tag{4.11}$$

$$diversity\_penalty = \left||mean\_pred\_axis\_0 - mean\_pred\_axis\_1\,|\right| * (1 - mean(attention\_score)) \tag{4.12}$$

$$temporal\_reg = e^{-mean(y_{pred})} \tag{4.13}$$

$$avg\_attention = \sqrt{mean(attention\_score)} \tag{4.14}$$

Where $pred\_correct$ is the cross entropy between the true values and the predicted values and the mathematical representation is given by the equation (4.8). The $diversity\_penalty$ is different from other equations by including the mean of the attention score. It will penalize if the attention score is not distributed properly and it is computed by the equation (4.12). The term $mean\_pred\_axis\_0$ is computed by the mean of the temporal predictions of the emotions and $mean\_pred\_axis\_1$ is the mean of probabilities of each emotions in a batch. The attention score is calculated using the equation (4.1). The $avg\_attention$ represents the square root of average attention score and it is mathematically represented by equation (4.14), the use of square root and logarithmic transformation introduces non-linearities and scaling effects that can influence the contribution of each component. The term $temporal\_reg$ represents the temporal regularization which influences the model to have lower temporal influence when the mean predicted values are high. It is computed using the equation (4.13) which is influenced by y_{pred}. This helps prevent the model from being overly influenced by strong temporal patterns, potentially reducing overfitting.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 PREPROCESSED DATA

The frames are extracted from the videos and then they undergo transformation in shape, sharpness of the image to maintain the consistency of the dataset. Colour normalization will be applied to mitigate variations in lighting conditions, promoting robustness in feature extraction. These processing technique is applied to the Figure 5.1 and get the processed image in Figure 5.2.

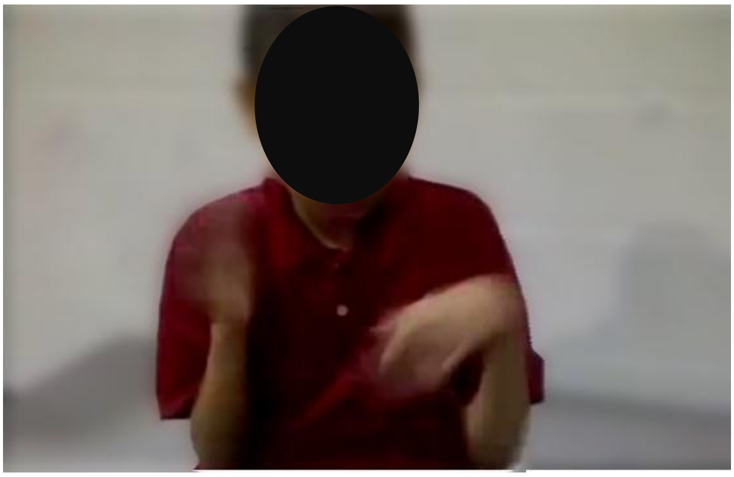


Figure 5.1 – Frame without preprocessing

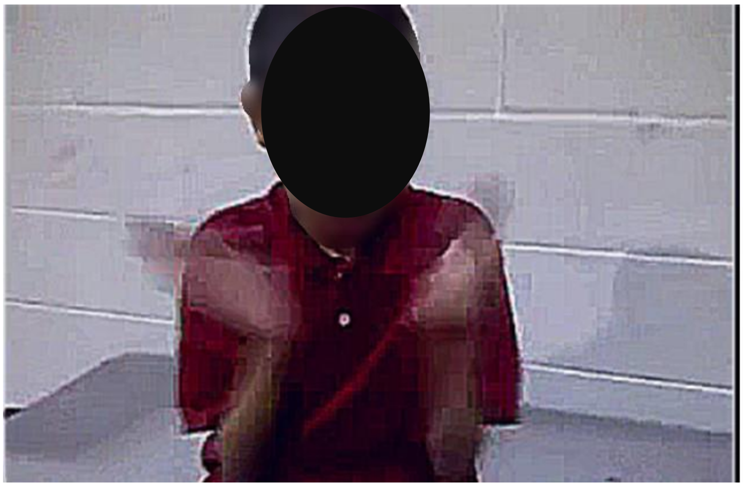


Figure 5.2 – Frame after preprocessing

5.2 LANDARK POINTS EXTRACTION:

Landmark point extraction from children involves capturing detailed information about facial, hand, and body features. This comprehensive set of 468 face landmarks, along with 21 hand and 33 body landmarks, enables a thorough representation of expressive behaviours. Facial landmarks, intricately mapping facial features, contribute to nuanced emotion analysis. Simultaneously, hand and body landmarks provide insights into gestural and postural aspects, enhancing the model's understanding of expressive cues. If the hands of the children is not visible in the frame, then they won’t be considered for the emotion detection. Figure 5.3 represents the visualization of the landmark points in the face and body and the Figure 5.4 represents the facial coordinates of the Figure 5.3 and the Figure 5.5 represents the right hand coordinate points of the Figure 5.3.

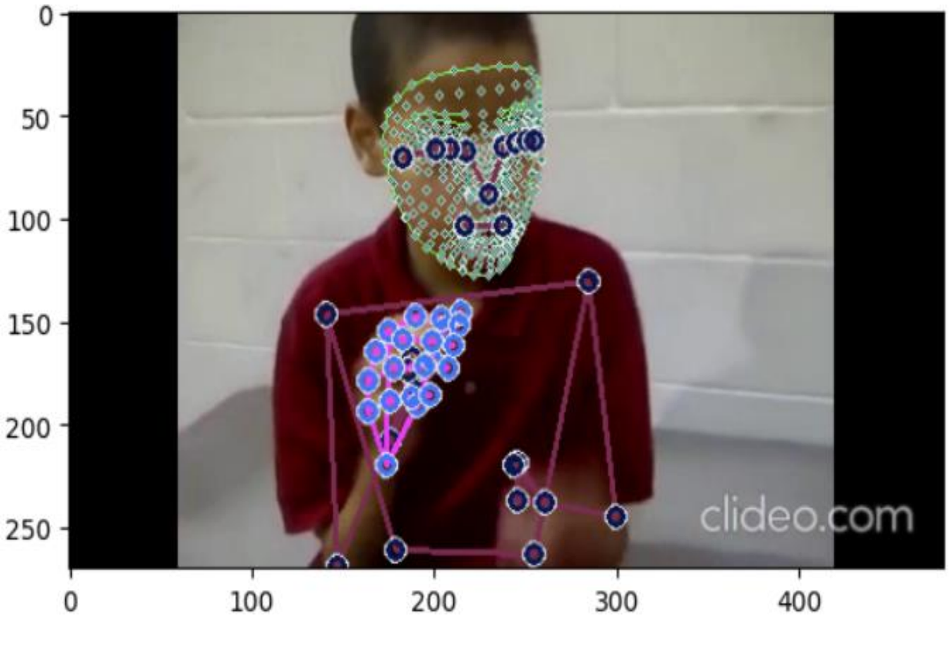
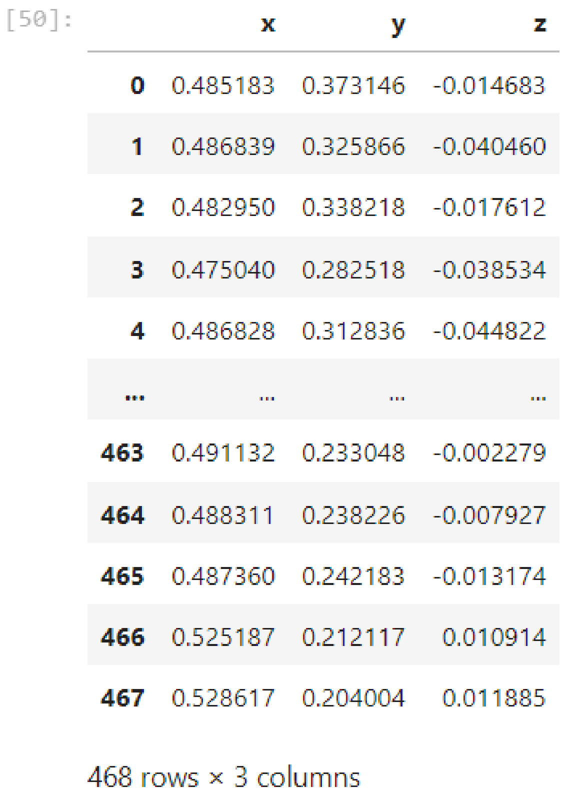


Figure 5.3 – Visualization of Landmark points of the body

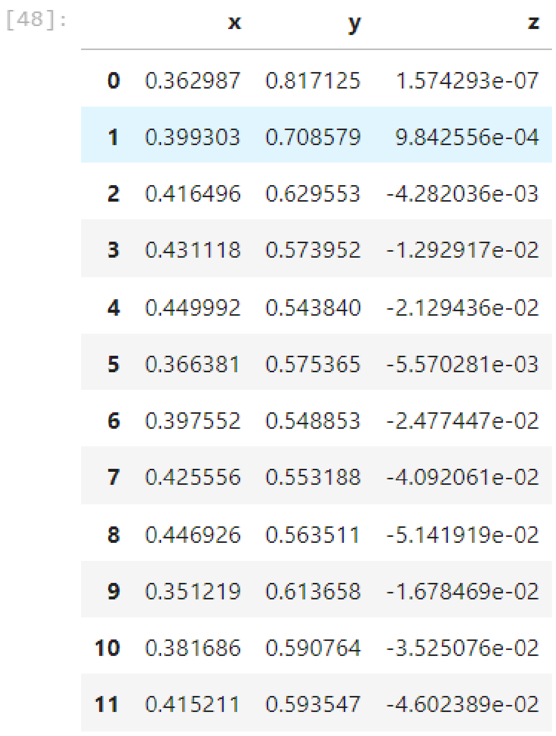


[50]:

	x	y	z
0	0.485183	0.373146	-0.014683
1	0.486839	0.325866	-0.040460
2	0.482950	0.338218	-0.017612
3	0.475040	0.282518	-0.038534
4	0.486828	0.312836	-0.044822
...	...	...	...
463	0.491132	0.233048	-0.002279
464	0.488311	0.238226	-0.007927
465	0.487360	0.242183	-0.013174
466	0.525187	0.212117	0.010914
467	0.528617	0.204004	0.011885

468 rows × 3 columns

Figure 5.4 – Coordinates of Facial landmark points



[48]:

	x	y	z
0	0.362987	0.817125	1.574293e-07
1	0.399303	0.708579	9.842556e-04
2	0.416496	0.629553	-4.282036e-03
3	0.431118	0.573952	-1.292917e-02
4	0.449992	0.543840	-2.129436e-02
5	0.366381	0.575365	-5.570281e-03
6	0.397552	0.548853	-2.477447e-02
7	0.425556	0.553188	-4.092061e-02
8	0.446926	0.563511	-5.141919e-02
9	0.351219	0.613658	-1.678469e-02
10	0.381686	0.590764	-3.525076e-02
11	0.415211	0.593547	-4.602389e-02

Figure 5.5 – Coordinates of right hand landmark points

5.3 EMOTION PREDICTION:

At the time of predicting the emotion, the model gives the probabilities of each emotion in each frame. As the children are autistic in nature, it will be beneficial to consider the top two emotions with highest probabilities. Figure 5.6 represents the intermediate frame of the predicted video with labelled landmark points.

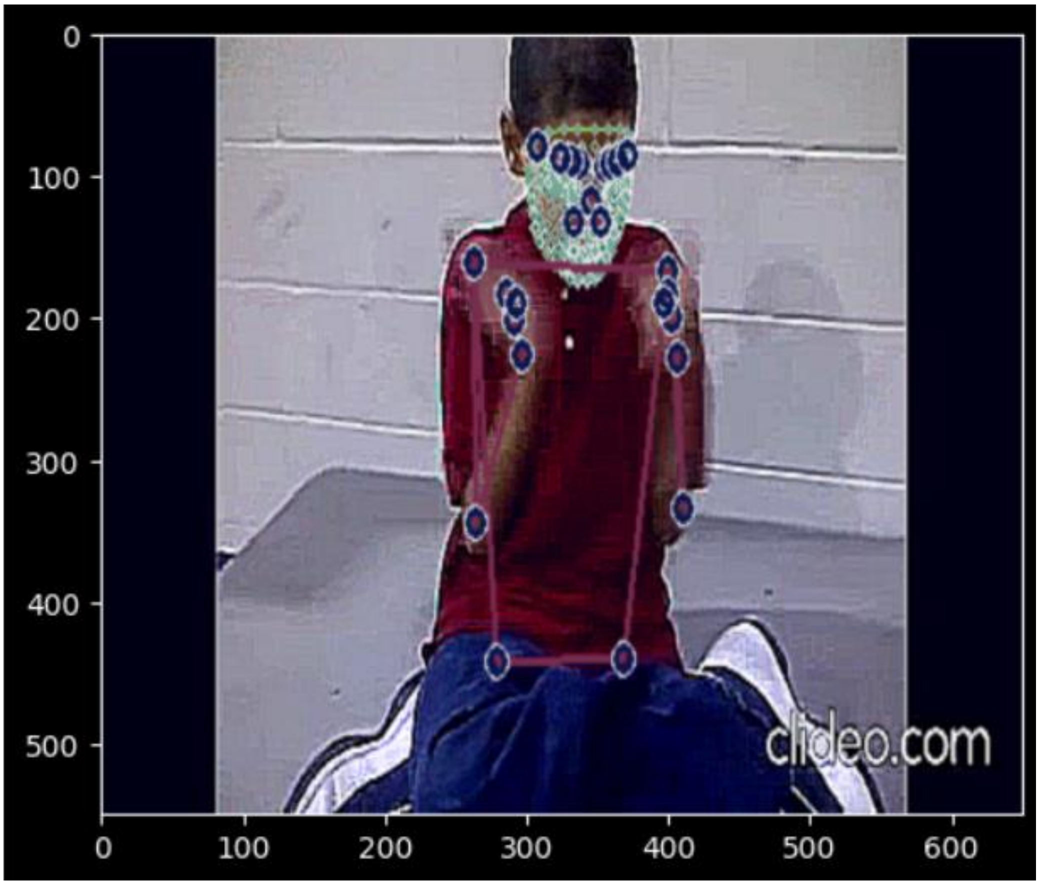
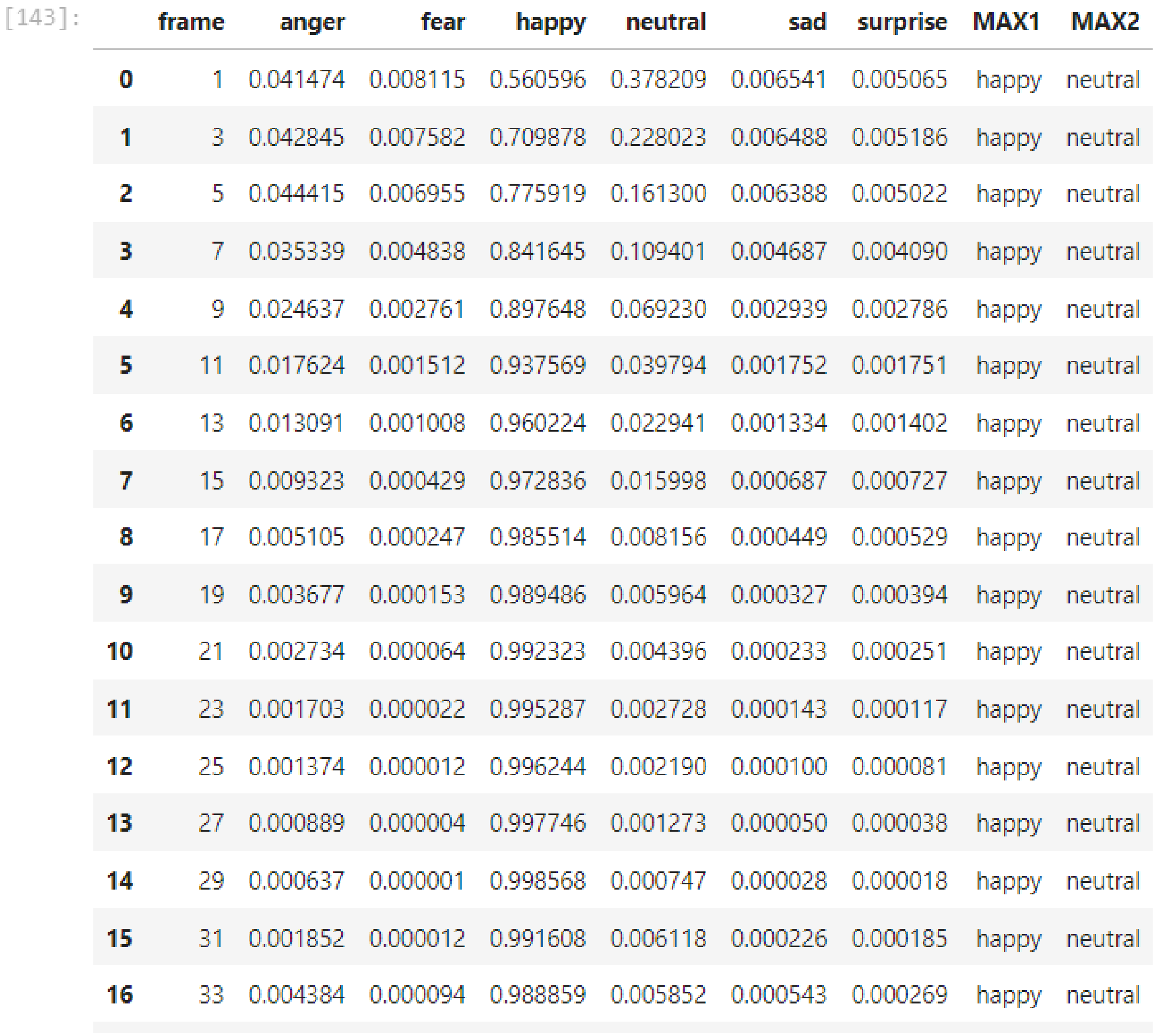


Figure 5.6 – Intermediate frame of the predicted video

The subsequent image is structured as a tabular representation which is shown in the Figure 5.7, with each row corresponds to a distinct frame. The 'frame' column serves as the index of the alternate frames in the predicted video, while the subsequent columns, such as 'anger,' 'fear,' 'happy,' 'neutral,' 'sad,' and 'surprise,' contain predicted probabilities for the respective emotion categories. Additionally, the 'MAX1' and 'MAX2' columns identify the two emotion categories with the highest scores for each frame.



[143]:

	frame	anger	fear	happy	neutral	sad	surprise	MAX1	MAX2
0	1	0.041474	0.008115	0.560596	0.378209	0.006541	0.005065	happy	neutral
1	3	0.042845	0.007582	0.709878	0.228023	0.006488	0.005186	happy	neutral
2	5	0.044415	0.006955	0.775919	0.161300	0.006388	0.005022	happy	neutral
3	7	0.035339	0.004838	0.841645	0.109401	0.004687	0.004090	happy	neutral
4	9	0.024637	0.002761	0.897648	0.069230	0.002939	0.002786	happy	neutral
5	11	0.017624	0.001512	0.937569	0.039794	0.001752	0.001751	happy	neutral
6	13	0.013091	0.001008	0.960224	0.022941	0.001334	0.001402	happy	neutral
7	15	0.009323	0.000429	0.972836	0.015998	0.000687	0.000727	happy	neutral
8	17	0.005105	0.000247	0.985514	0.008156	0.000449	0.000529	happy	neutral
9	19	0.003677	0.000153	0.989486	0.005964	0.000327	0.000394	happy	neutral
10	21	0.002734	0.000064	0.992323	0.004396	0.000233	0.000251	happy	neutral
11	23	0.001703	0.000022	0.995287	0.002728	0.000143	0.000117	happy	neutral
12	25	0.001374	0.000012	0.996244	0.002190	0.000100	0.000081	happy	neutral
13	27	0.000889	0.000004	0.997746	0.001273	0.000050	0.000038	happy	neutral
14	29	0.000637	0.000001	0.998568	0.000747	0.000028	0.000018	happy	neutral
15	31	0.001852	0.000012	0.991608	0.006118	0.000226	0.000185	happy	neutral
16	33	0.004384	0.000094	0.988859	0.005852	0.000543	0.000269	happy	neutral

Figure 5.7 – Probabilities of emotions in each frame

5.4 PERFORMANCE COMPARISON BETWEEN REWARD FUNCTION

Now, the performance of the prediction without reward function and with different reward functions. The graphical representation of the training loss and training accuracy for all the functions are shown in the Figure 5.8, Figure 5.9, Figure 5.10, Figure 5.11. They show how the training loss and training accuracy changes in each epochs

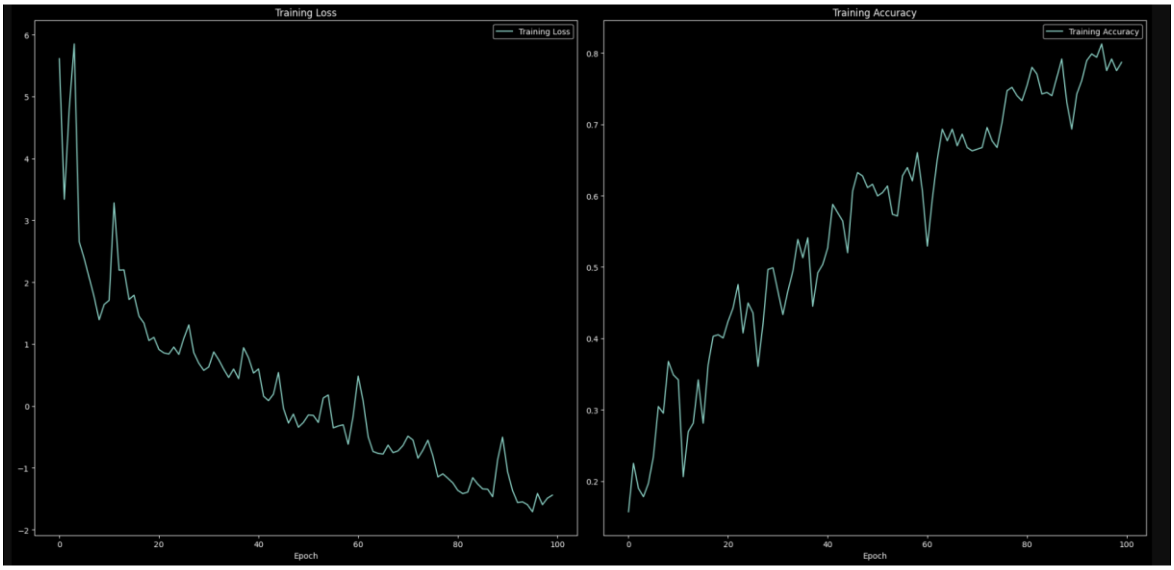


Figure 5.8 – Training loss and training accuracy graph with reward function 1

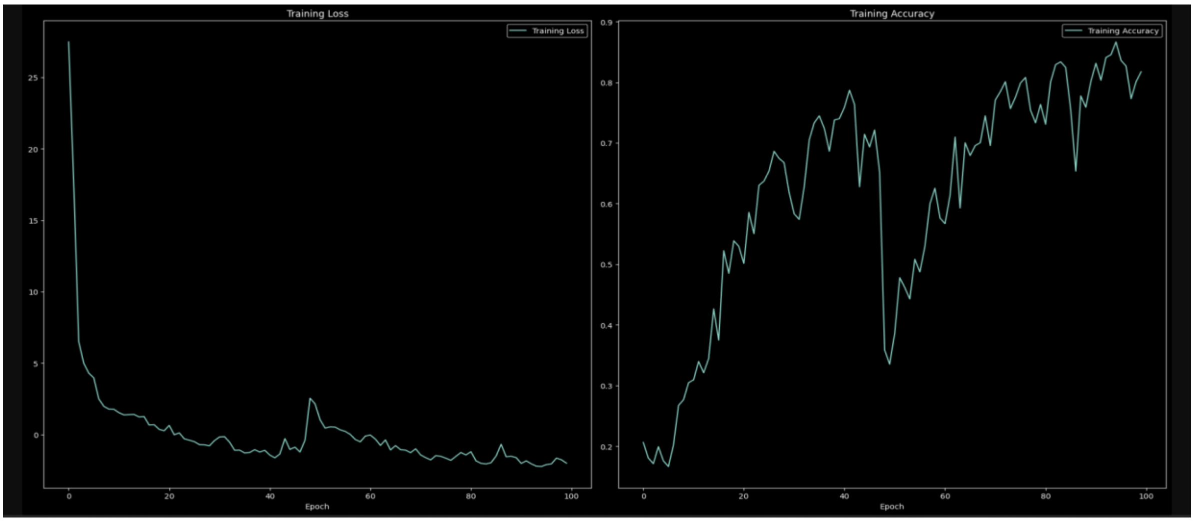


Figure 5.9 – Training loss and training accuracy graph with reward function 2

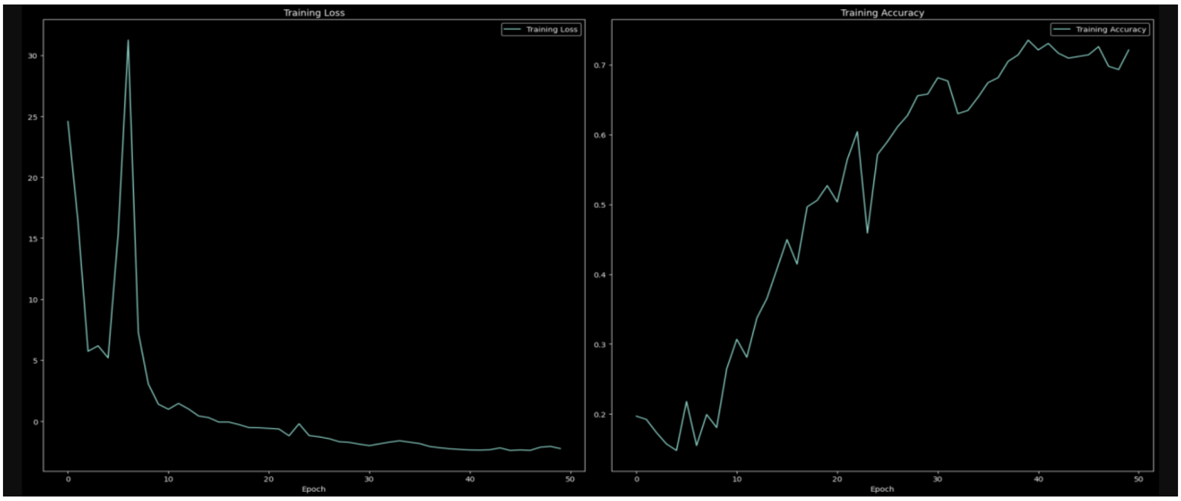


Figure 5.10 – Training loss and training accuracy graph with reward function 3

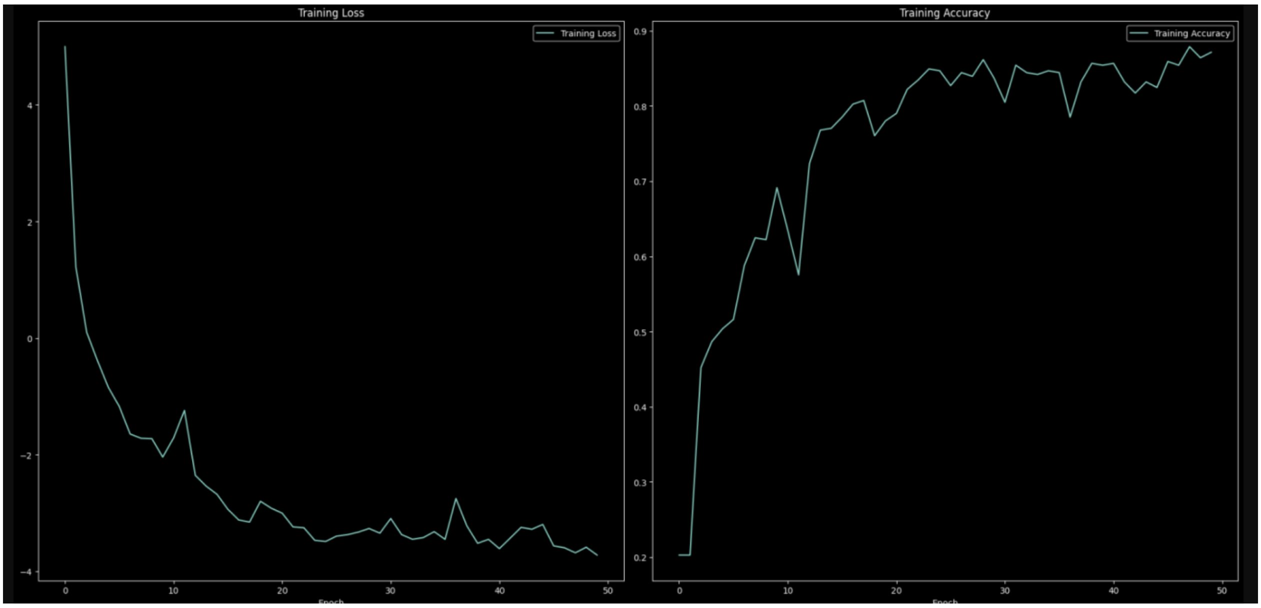


Figure 5.11 – Training loss and training accuracy graph with reward function 4

Then these models with different reward functions are tested with the test data which give the different accuracies for each function, some function gives good accuracy and some functions yields lesser accuracy than the model without any reward function.

Table 5.1 – Performance comparison based on the accuracy score

FUNCTION	**ACCURACY**
No Reward function	79.26
Reward Function 1	78.52
Reward Function 2	82.96
Reward Function 3	77.78
Reward Function 4	88.15

Table 5.1 contains the function along with their respective accuracy for the test data prediction and the accuracies of each functions is shown in a graphical representation in Figure 5.12.

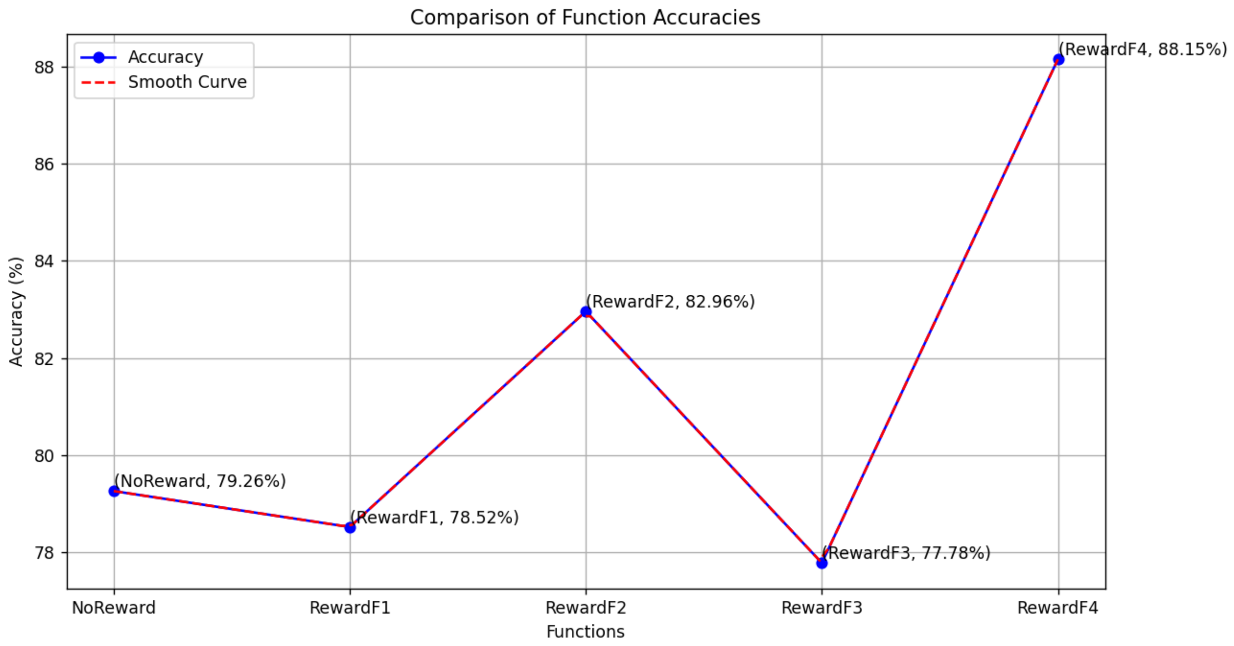


Figure 5.12 – Graphical representation of performance of the different reward functions

The performance evaluation of different reward functions shows different impacts on the accuracy of the model. In the absence of a specific reward function that only attention mechanism is implemented which achieves an accuracy of 79.26%. After introducing the first reward function, leads to a slightly lower accuracy of 78.52%, suggesting that this particular reward function might not significantly enhance the model's performance. However, second reward function shows a notable improvement with an accuracy of 82.96%, which indicates its positive impact on predictive capabilities on the test data. But the third reward function yields an accuracy of 77.78%, slightly below the baseline. The most promising result comes from Reward Function 4, which achieves the highest accuracy of 88.15%, making it the most effective in enhancing the model's overall performance. The selection of a proper reward function is important for optimizing the model's performance, and Reward Function 4 shows better performance in this comparative analysis.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

The project involves in the development of a specialized emotion recognition system tailored to the complex expressive behaviours exhibited by autistic children. The project will focus on exploring facial expressions and body movements as key modalities for emotion recognition. A significant aspect of the project's scope involves addressing the challenge of occluded faces by incorporating methods for face reconstruction when parts of the face are hidden and other challenges like lack of diverse labelled dataset, developing the proper reward function for the diverse emotion. An LSTM layer captures temporal patterns in frame sequences, supports dynamic expression comprehension. Attention Mapping and Reward Function enhance focus and offer positive reinforcement during training.

The reward function in the proposed architecture is an important element designed to guide the learning process effectively. It is composed with several key factors that collectively contribute to enhancing the model's performance. Mainly, the reward function majorly relies on the model's ability to make correct predictions. Accurate emotion recognition is encouraged and reinforced through positive reward. Additionally, the function incorporates a diversity penalty to discourage the model from consistently predicting the same emotion. The average attention scores from the attention mechanism play a pivotal role, influencing the reward and emphasizing the significance of selectively attending to relevant features. Collectively, these components form a comprehensive reward framework, steering the model towards improved accuracy, robustness, and adaptability in the intricate task of emotion recognition. The Fully Connected Layer integrates attention information. The reward during loss helps learning through positive reinforcement. The output layer predicts six emotions with probability distributions. The highest probability class is the predicted emotion.

6.2 FUTURE WORK

The proposed work’s performance can be improved by optimization of the model, a process that can be achieved through the exploration of advanced optimization techniques and fine-tuning of hyperparameters. Collaborative efforts with clinicians are crucial for the validation of the model's predictions against clinical assessments, ensuring the accuracy and reliability of the emotion recognition system. Addressing privacy concerns is necessary for the implementation of privacy measures to protect sensitive data. Challenges in the project, such as irregular face alignment in frames and limited visibility of body parts, can be mitigated. Facial derotation techniques can be implemented to address irregular face alignment, ensuring that the capture of facial expressions even in challenging scenarios are precise. Additionally, advanced face recovery techniques can be employed to enhance the reconstruction of occluded or partially visible facial features, further improving the system's overall performance. These techniques contribute to a adaptable emotion recognition system for Autism Spectrum Disorder (ASD) children.

REFERENCES

[1] Babu P. R. K 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.

[2] 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.

[3] Hashemi J et al., “Computer vision analysis for quantification of autism risk behaviors,” IEEE Trans. Affect. Comput., vol. 12, no. 1, pp. 215–226, First Quarter 2021

[4] Li Z, X. Deng, X. Li, and L. Yin, “Integrating semantic and temporal relationships in facial action unit detection,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 5519–5527

[5] Nie G, Ullal A, Zheng Z, Swanson AR, Weitlauf AS, Warren ZE, Sarkar N. An Immersive Computer-Mediated Caregiver-Child Interaction System for Young Children With Autism Spectrum Disorder. IEEE Trans Neural Syst Rehabil Eng. 2021;29:884-893. doi: 10.1109/TNSRE.2021.3077480. Epub 2021 May 18. PMID: 33945481; PMCID: PMC8189254.

[6] Omar K. S, P. Mondal, N. S. Khan, M. R. K. Rizvi, and M. N. Islam, “A machine learning approach to predict autism spectrum disorder,” in 2019 International conference on electrical, computer and communication engineering (ECCE). IEEE, 2019, pp. 1–6.

[7] Pons G 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.

[8] 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.

[9] Shao Z, 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.

[10] Stef van der Struijk, Hung-Hsuan Huang, Maryam Sadat Mirzaei, Toyoaki Nishida FACSvatar: An Open Source Modular Framework for Real-Time FACS based Facial Animation. IVA '18: Proceedings of the 18th International Conference on Intelligent Virtual Agents.

[11] Takc H ¸ı and S. Yes¸ilyurt, “Diagnosing autism spectrum disorder using machine learning techniques,” in 2021 6th international conference on computer science and engineering (ubmk). IEEE, 2021, pp. 276–280.

[12] Talaat M, 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.

[13] Tellamekala M. K, 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.

[14] Vakadkar ,D. Purkayastha, and D. Krishnan, “Detection of autism spectrum disorder in children using machine learning techniques,” SN Computer Science, vol. 2, pp. 1–9, 2021.

[15] Yan, J. Wang, Q. Li, C. Wang, and S. Pu, “Self-supervised regional and temporal auxiliary tasks for facial action unit recognition,” in Proc. ACM Int. Conf. Multimedia, 2021, pp. 1038–1046.

[16] 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).

[17] 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.

[18] Zhao et al., “Atypical head movement during face-to-face interaction in children with autism spectrum disorder,” Autism Res., vol. 14, no. 6, pp. 1197–1208, 2021