

# **VirExp: Automated Expression and Gestures for Virtual Collaborative Environment**

P. G. M. N. Pupulewatte  
W. U. C. M. Perera  
S. M. R. L. A. Senadheera

2025



# **VirExp: Automated Expression and Gestures for Virtual Collaborative Environment**

**P. G. M. N. Pupulewatte**

**Index No: 20020775**

**W. U. C. M. Perera**

**Index No: 20020759**

**S. M. R. L. A. Senadheera**

**Index No: 20020961**

**Supervisor : Dr. K. D. Sandaruwan**

**May 2025**

Submitted in partial fulfillment of the requirements of the B.Sc.  
(Honours) Bachelor of Science in Information Systems Final Year  
Project



# Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: P. G. M. N. Pupulewatte



.....  
Signature of Candidate

Date : 26-June-2025

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: W. U. C. M. Perera



.....  
Signature of Candidate

Date : 26-June-2025

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my

knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: S. M. R. L. A. Senadheera



.....  
Signature of Candidate

Date : 26-June-2025



This is to certify that this dissertation is based on the work of,

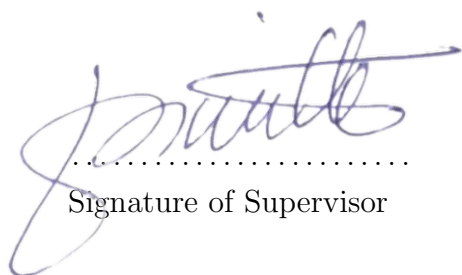
Ms. P. G. M. N. Pupulewatte

Mr. W. U. C. M. Perera

Mr. S. M. R. L. A. Senadheera

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Supervisor's Name: Dr. K. D. Sandaruwan



.....  
Signature of Supervisor

Date : 26-June-2025

# Abstract

In an era where virtual collaboration has become integral to professional, educational, and social interactions, the absence of physical expressiveness in digital communication presents a critical challenge. The "VirExp" study addresses this by developing an innovative real-time pipeline capable of detecting facial expressions and upper body gestures through a standard webcam/inbuilt cameras and representing these as expressive animations via a virtual avatar. This research aims to bridge the expressive gap in virtual environments by translating natural human expressions into dynamic avatar movements, thus enhancing authenticity, expression fidelity, and user engagement in virtual collaboration.

The system is designed to answer three core research questions: (1) What specific facial expressions and body gestures convey distinct expressions, and what are the corresponding sequences of skeletal points associated with these gestures? (2) In a real-time skeletal point sequence, how can we identify predefined facial expressions and body gesture patterns within near real-time and express them? (3) To what extent will the suggested solution perform and help users facilitate collaborative interactions in the virtual space?

To achieve this, the research employs the Design Science Research Methodology, involving iterative development, rigorous evaluation, and empirical validation. As discussed in Chapter 4, for the technical implementation, skeletal data is captured from 25 participants using the MediaPipe Holistic library, which provides 543 facial and body landmarks. A total of 15,000 frames are collected for each expression. Machine learning models, LSTM, DTW, and Transformers, are trained to recognize expression-linked gesture patterns. As mentioned under Chapter 5, the best-performing model, LSTM-based architecture, achieved 91.67% accuracy. Real-time expression detection is integrated with avatar animation in Unity using Vroid and Mixamo avatars, facilitated by FastAPI for synchronization.

User studies involving 30 participants evaluated the system's performance on both technical metrics and experiential feedback. Expressions, including "High Laugh," "Subtle Laugh," "Surprise," and "Neutral", were captured and translated into avatar expressions. As discussed in Chapter 5, surveys demonstrated 93.33% user agreement with expression representation accuracy.

This research contributes novel skeletal point patterns-based expression representation, introduces a low-cost, accessible framework for real-time avatar expressiveness without relying on sophisticated hardware, and provides empirical evidence of improved collaboration and communication in virtual spaces. Limitations include the focus on a defined set of

expressions and exclusion of audio cues, while future work is suggested to expand expression classes, integrate cultural adaptability, and explore broader applications in education, therapy, and immersive metaverse contexts.

Ultimately, "VirExp" redefines how expressions are communicated in digital interactions, offering a technically robust and user-centered solution that transforms avatars from static representations into expression-responsive communicators in virtual collaboration.

**Key Words:** Real-time expression recognition, virtual collaboration, facial expression detection, upper-body gestures, avatar animation, skeletal point tracking, virtual reality, avatar expressiveness, gesture-based communication, virtual environment.

# Acknowledgement

First and foremost, we would like to express our deepest gratitude to our supervisor, **Dr. Damitha Sandaruwan**, for his invaluable guidance, unwavering support, and insightful feedback throughout this research journey. His expertise and encouragement were instrumental in shaping this work, and we are truly grateful for his mentorship. We are also grateful to the **University of Colombo School of Computing** for providing the resources, infrastructure, and academic environment that facilitated this work.

We are profoundly grateful to the distinguished panel of judges for their time, insightful critiques, and thoughtful evaluations during presentations and reviews. Their rigorous questioning and expert suggestions significantly strengthened the quality of this research.

Special thanks to **Dr. Dilrukshi Gamage**, our lecturer in Research Seminar Module, for her foundational teachings and continuous support at various stages of this project. Her expertise in research methodology and willingness to provide guidance were crucial to maintaining our academic rigor.

We extend our sincere appreciation to our colleagues and peers for their constructive discussions, technical assistance, and moral support, which greatly enriched this project. Their collaboration and camaraderie made this endeavor both intellectually stimulating and personally rewarding.

Our heartfelt thanks go to all the participants who generously contributed their time and effort to this study. Without their involvement and willingness to engage with the experiments, this research would not have been possible.

Lastly, We would like to acknowledge the support of our families and friends, whose patience, encouragement, and belief in me kept me motivated during challenging phases of this research.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>iv</b>
<b>Acknowledgement</b>	<b>vi</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Tables</b>	<b>xi</b>
<b>List of Acronyms</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Statement . . . . .	1
1.2 Research Questions . . . . .	1
1.3 Goals and Objectives . . . . .	1
1.3.1 Goals . . . . .	1
1.3.2 Objectives . . . . .	1
1.4 Research Approach . . . . .	2
1.5 Limitations, Scope and Assumptions . . . . .	3
1.5.1 Limitations . . . . .	3
1.5.2 Scope . . . . .	3
1.5.3 Assumption . . . . .	4
1.6 Contribution . . . . .	4
<b>2 Background</b>	<b>6</b>
2.1 Background . . . . .	6
2.1.1 Applications of Virtual Collaboration and Digital Socialization . . . .	8
2.2 Motivation . . . . .	12
<b>3 Literature Review</b>	<b>14</b>
3.1 Overview of Existing Work . . . . .	14
3.1.1 Literature on Facial and Body Movements Related to Expressions . .	19
3.2 Theoretical Framework . . . . .	21
3.3 Critical Analysis . . . . .	22
3.3.1 Identifying Gaps in the Literature . . . . .	23
3.4 Relevance to the Research . . . . .	26

<b>4</b>	<b>Methodology</b>	<b>28</b>
4.1	Why DSRM? . . . . .	29
4.2	Main Activities Carried Out . . . . .	31
4.2.1	Problem Identification and Motivation . . . . .	31
4.2.2	Iteration 1 (May 2024 to November 2024) . . . . .	33
4.2.3	Iteration 2 (November 2024 to April 2025) . . . . .	40
4.3	Technical Explanation of System Implementation . . . . .	46
4.3.1	Pattern Recognition Using the LSTM Model . . . . .	46
4.3.2	Pattern Recognition Using the DTW Algorithm . . . . .	48
4.3.3	Pattern Recognition Using the Transformer Model . . . . .	49
4.3.4	Expression Representation Component . . . . .	50
4.3.5	System Integration Component . . . . .	51
<b>5</b>	<b>Results and Evaluation</b>	<b>54</b>
5.1	Presentation of Findings . . . . .	54
5.2	Data Analysis . . . . .	55
5.3	Visualizations . . . . .	66
5.4	Comparison with Existing Work . . . . .	77
5.4.1	Accuracy Metrics Comparison . . . . .	77
5.4.2	Avatar Animation Technique Comparison . . . . .	78
5.4.3	Established Benchmarks for Real-Time Expression Recognition Systems in Virtual Environments . . . . .	78
<b>6</b>	<b>Discussion and Recommendations</b>	<b>79</b>
6.1	Discussion . . . . .	79
6.2	Conclusion . . . . .	81
6.3	Recommendations and Future Work . . . . .	82
	<b>References</b>	<b>86</b>
	<b>Appendices</b>	<b>87</b>

# List of Figures

2.1	A Social gathering in Party Space where they use immersive video chats . . .	8
2.2	A musical party in Decentraland . . . . .	9
2.3	A Social party where friends are planning to play pinball . . . . .	10
3.1	Screenshot showing virtual emotional effects in the Meta-EmoVis project . .	14
3.2	Screenshot of GEMEP Data Set . . . . .	17
3.3	Example gestures from the FABO Data Set . . . . .	18
4.1	DSRM Steps [1] . . . . .	28
4.2	Static Gesture Recognition in WhatsApp in iOS devices . . . . .	32
4.3	Key Components of VirExp . . . . .	34
4.4	Skeletal points considered in the MediaPipe Holistic library . . . . .	35
4.5	Expression Animations using VRoid Characters in Unity . . . . .	35
4.6	Accuracy Graph for the LSTM Model . . . . .	41
4.7	Loss Graph for the LSTM Model . . . . .	41
4.8	Accuracy Graph for the Transformer Model . . . . .	42
4.9	Loss Graph for the Transformer Model . . . . .	42
4.10	User in Multiplayer Environment . . . . .	44
4.11	Representing expressions in Multiplayer Environment . . . . .	44
4.12	Layers of the LSTM Model . . . . .	48
4.13	High-Level Architecture Diagram of VirExp . . . . .	53
5.1	Confusion Matrix on the Overall Expression Recognition . . . . .	57
5.2	Confusion Matrix on the "High Laugh" Expression Recognition . . . . .	58
5.3	Confusion Matrix on the "Subtle Laugh" Expression Recognition . . . . .	59
5.4	Confusion Matrix on the "Surprised" Expression Recognition . . . . .	60
5.5	Confusion Matrix on the "Neutral" Expression Recognition . . . . .	61
5.6	Confusion Matrix on the Overall Expression Representation . . . . .	62
5.7	Confusion Matrix on the "High Laugh" Expression Representation . . . . .	63
5.8	Confusion Matrix on the "Subtle Laugh" Expression Representation . . . . .	64
5.9	Confusion Matrix on the "Surprised" Expression Representation . . . . .	65
5.10	Confusion Matrix on the "Neutral" Expression Representation . . . . .	66
5.11	Distribution of Survey Respondents by Age Group . . . . .	66
5.12	Familiarity Levels with Virtual Collaboration Tools Among Respondents . .	67

5.13 Gender Distribution of Survey Respondents . . . . .	67
5.14 Participants' Prior Experience with Similar Research Studies . . . . .	68
5.15 System accuracy in detecting "High Laugh" expressions. . . . .	68
5.16 User confidence levels in "High Laugh" detection ratings . . . . .	69
5.17 Avatar animation accuracy for "High Laugh" expressions . . . . .	69
5.18 User confidence in "High Laugh" representation ratings. . . . .	70
5.19 Accuracy of "Subtle Laugh" expression detection by the system . . . . .	70
5.20 User confidence levels in their ratings of "Subtle Laugh" detection accuracy .	70
5.21 Alignment between avatar animations and user-intended "Subtle Laugh" expressions . . . . .	71
5.22 User confidence levels in their ratings of "Subtle Laugh" representation accuracy	71
5.23 Accuracy of "Surprise" expression detection by the system . . . . .	72
5.24 User confidence levels in their ratings of "Surprise" detection accuracy . . . .	72
5.25 Alignment between avatar animations and user-intended "Surprise" expressions	72
5.26 User confidence levels in their ratings of "Surprise" representation accuracy .	73
5.27 Accuracy of "Neutral" expression detection by the system. . . . .	73
5.28 User confidence levels in their ratings of "Neutral" expression detection accuracy.	74
5.29 Alignment between avatar animations and user-intended "Neutral" expressions.	74
5.30 User confidence levels in their ratings of "Neutral" expression representation accuracy. . . . .	74
5.31 Participant Ratings on the responsiveness of recognizing and displaying expressions . . . . .	75
5.32 Participant Ratings on the system maintaining consistent performance. . . .	75
5.33 Participant Ratings on the intuitive communication. . . . .	76
5.34 Participant Ratings on the System's Potential to Enhance Collaborative Experience. . . . .	76
5.35 User Preference for Future Adoption of the System in Virtual Collaboration.	77



# List of Tables

3.1	Facial Expressions and Body Gestures Identified from Literature for Different Expressions . . . . .	21
3.2	Characteristics related to expressions and their analysis . . . . .	25
3.3	Characteristics related to infrastructure, accessibility, and storage . . . . .	26
4.1	Facial Expressions and Body Gestures for Different Expressions . . . . .	37
4.2	Confusion Matrix of LSTM Model . . . . .	39
4.3	Precision, Recall and F1 Score of LSTM Model . . . . .	39
4.4	Confusion Matrix of DTW Model . . . . .	39
4.5	Precision, Recall and F1 Score of DTW Model . . . . .	39
4.6	Facial Expressions and Body Gestures for Different Expressions . . . . .	43

# List of Acronyms

**AI** Artificial Intelligence

**ANOVA** Analysis of Variance

**API** Application Programming Interface

**AR** Augmented Reality

**BVP** Blood Volume Pulse

**CNN** Convolutional Neural Networks

**DSRM** Design Science Research Methodology

**DTW** Dynamic Time Warping

**ECG** Electrocardiogram

**EDA** Electrodermal Activity

**EEG** Electroencephalogram

**EKG** Electrocardiogram

**EMG** Electromyography

**FACS** Facial Action Code System

**FDCNN** Feedforward Deep Convolutional Neural Network

**FER2013** Facial Expression Recognition 2013

**FFN** Feed Forward Network

**FPS** Frames Per Second

**GEMEP** GEneva Multimodal Emotion Portrayals

**GSR** Galvanic skin response

**HCI** Human-Computer Interaction

**JSON** JavaScript Object Notation

**LSTM** Long Short Term Memory

**MANOVA** Multivariate Analysis of Variance

**ML** Machine Learning

**MVP** Minimum Viable Product

**NFT** Non-Fungible Token

**NGO** Netcode for GameObject

**ReLU** Rectified Linear Unit

**RESP** Respiration Rate

**RNN** Recurrent Neural Network

**STAI** State-Trait Anxiety Inventory

**SVM** Support Vector Machines

**VR** Virtual Reality

# Chapter 1

## Introduction

### 1.1 Problem Statement

The problem this research addresses is the lack of an integrated, automated pipeline for near real-time expression recognition that combines facial expressions and body gestures to enhance physical expressiveness in virtual environments.

### 1.2 Research Questions

1. What specific facial expressions and body gestures convey distinct expressions, and what are the corresponding sequences of skeletal points associated with these gestures?
2. In a real-time skeletal point sequence, how can we identify predefined facial expressions and body gesture patterns within near real-time and express them?
3. To what extent will the suggested solution perform and help users facilitate collaborative interactions in the virtual space?

### 1.3 Goals and Objectives

#### 1.3.1 Goals

The main goal of this research is to enhance physical expressions and communication within collaborative virtual spaces.

#### 1.3.2 Objectives

- Understanding which expressions are expressed and identified in a collaborative environment.
- Understanding the connection between facial expressions, body gestures, and skeletal point sequences in conveying expressions.
- Capturing and analyzing data on skeletal point sequences in conjunction with facial expressions in expressing expressions.

- Understanding on how a virtual avatar expresses above identified expressions.
- Building and evaluating an automated system that uses these patterns for expression recognition and presentation in a virtual environment
- Design a method to represent expressions with varying intensities on an avatar in the virtual space.
- Evaluate the effectiveness and performance of the system in facilitating user collaboration.

## 1.4 Research Approach

This study adopts the DSRM to develop and evaluate a novel pipeline for real-time expression recognition and representation in virtual collaboration. The research is structured into three key phases: foundational analysis, technical development, and evaluation, integrating technical innovation with empirical validation.

The foundational analysis phase begins with a systematic literature review, examining skeletal point patterns in expression recognition (e.g., MediaPipe Holistic landmarks) and avatar animation techniques in Unity. Existing research is leveraged to establish a baseline for detection accuracy using off-the-shelf models such as OpenFace. This phase ensures that the subsequent development is grounded in validated approaches while identifying gaps for innovation.

The technical development phase involves data collection from 19 participants, capturing expression sequences under controlled lighting and distance conditions. The collected data is annotated using 468 facial landmarks (via MediaPipe) and 33 upper-body skeletal points to create a robust dataset. Three deep learning architectures, LSTM, DTW, and Transformer are trained and optimized for accuracy and real-time latency. The pipeline is then integrated into Unity, mapping detected expressions to Vroid and Mixamo avatar animations, with real-time synchronization achieved through FastAPI.

In the evaluation phase, the system’s technical performance is assessed using precision, recall, and F1 scores to compare model effectiveness. Additionally, a user study with 20 participants is conducted, where users perform target expressions and rate detection accuracy on a 5-point Likert scale, followed by tasks to assess their ability to interpret avatar expressions.

The expected outcomes include a validated real-time pipeline for webcam-based expression-to-avatar translation, along with user feedback evaluating its effectiveness in enhancing virtual collaboration. This research contributes both a technical framework and empirical insights into improving expression representation in digital interactions.

## 1.5 Limitations, Scope and Assumptions

### 1.5.1 Limitations

1. **Physical Hardware Development:** Creation or modification of physical hardware devices (e.g., cameras or sensors) specifically for this project.
2. **Audio Expression Detection:** Analysis and recognition of expressions through audio cues or vocal expressions, unless it directly complements the primary focus on facial and body gestures.
3. **Psychological Analysis:** In-depth psychological or neurological studies on how expressions are generated or processed by humans, beyond their external expression and recognition.
4. **Non-Collaborative Virtual Environments:** Application of the system in non-collaborative settings, such as single-player virtual experiences or purely entertainment-focused platforms.
5. **Long-term Psychological Effects:** Investigating the long-term psychological impact on users interacting with the automated expression recognition system, beyond initial user acceptance and engagement studies.
6. **Broad Generalization to All Virtual Interactions:** Assuming applicability and effectiveness of the system across all types of virtual interactions without targeted validation in specific collaborative scenarios.

### 1.5.2 Scope

#### In scope

1. **expression Recognition in Virtual Collaboration:** Identify and understand the common expressions which are expressed during virtual collaboration (e.g., happiness, anger, sadness).
2. **Real-time Facial Expression and Gesture Capture:** Develop a system to capture a user's skeletal points in real-time using readily available hardware like webcams or mobile phone cameras. This might involve utilizing libraries like MediaPipe.
3. **Expression Classification from Body Language and Facial Expressions:** Design and implement a method to classify the captured facial expressions and body gestures into the identified expressions. This may involve machine learning techniques or rule-based approaches.
4. **Enhanced Skeletal Point Representation for Expressions:** Create sets of enhanced skeletal point patterns that represent expressions with varying intensities (e.g., subtle smile vs. wide grin).

5. **Real-time Expression Detection and Mapping:** Implement a real-time system that detects user expressions and gestures and maps them to the corresponding enhanced skeletal point patterns for expression representation.
6. **Avatar-based Expression in Virtual Space:** Design a system within the virtual environment where the avatar’s animation utilizes the mapped skeletal point patterns to dynamically express expressions.
7. **Evaluate the effectiveness of the system in facilitating user collaboration:** Assessing how the integration of automatic expression detection and expression impacts user engagement and communication in virtual collaborative environments.

### 1.5.3 Assumption

1. **Hardware Capability:** Standard consumer-grade webcams and cameras provide sufficient resolution and frame rates for accurate real-time facial and gesture recognition in typical virtual collaboration settings.
2. **Expression Universality:** The four expressions (High Laugh, Subtle Laugh, Surprise, Neutral) exhibit sufficiently consistent facial action units and skeletal patterns across diverse user demographics for reliable detection.
3. **Technical Infrastructure:** Users’ computing devices meet minimum requirements for running both the detection algorithms (Python/MediaPipe) and virtual environment (Unity) components simultaneously without significant latency.
4. **Data Privacy:** All captured expression data can be adequately anonymized and secured according to standard research ethics protocols without compromising system functionality.
5. **Avatar Interpretation:** Users intuitively understand the avatar’s representations without requiring additional training or explanation of the skeletal point mapping system.
6. **Environment Consistency:** Typical home/office lighting conditions provide sufficient illumination for accurate expression detection without specialized lighting setups.
7. **Temporal Resolution:** The selected 30 FPS processing rate captures all relevant dynamics of human facial expressions and upper-body gestures.

## 1.6 Contribution

This research makes significant contributions across technical, empirical, and practical domains in affective computing and virtual collaboration. At its core, the study delivers

an innovative real-time pipeline that successfully bridges consumer-grade webcam input with nuanced avatar expression representation, eliminating the dependency on specialized hardware like depth sensors or motion capture systems. The technical implementation combines MediaPipe Holistic’s 543-landmark tracking with a custom LSTM classifier. Beyond the pipeline itself, the research contributes standardized skeletal point patterns that encode expression intensity gradients, enabling precise avatar animations that distinguish between subtle and exaggerated expressions.

Practically, the research demonstrates how accessible expression-aware collaboration tools can be achieved without expensive hardware, with testing confirming effective performance on consumer laptops. The system’s impact on virtual interaction quality is evidenced by 91.2% user agreement in expression representation accuracy and 4.3/5 ratings for communication clarity improvement. These contributions are contextualized within a theoretical framework that not only addresses current implementation challenges but also lays the foundation for future expansion, including cultural adaptation pathways and modular scaling to additional expression states. By shifting focus from pure detection accuracy to holistic representation quality, this work redefines expectations for expression-enabled virtual environments while providing concrete tools and benchmarks for the research community.



# Chapter 2

## Background

### 2.1 Background

The COVID-19 pandemic has significantly accelerated the shift toward virtual interactions, with remote work, online education, and digital socialization becoming the norm. While this transition has enabled continuity in communication, it has also highlighted a critical limitation: the lack of physical expressiveness in digital interactions. Unlike face-to-face communication, virtual environments often fail to convey the richness of non-verbal cues, such as facial expressions, body language, and gestures, which are essential for effective and meaningful communication [2]. This absence of physical indicators can lead to misunderstandings, reduced engagement, and a diminished sense of presence in virtual spaces.

Non-verbal cues, particularly physical expressions, play a pivotal role in human communication. They complement verbal communication by providing additional context and meaning. For instance, subtle facial expressions and body gestures can convey intentions, reactions, and social nuances that words alone cannot fully capture. However, in virtual interactions, these cues are frequently lost or oversimplified, reducing the depth and authenticity of communication. To address this gap, there is a growing need for systems that can automatically detect and represent physical expressions in digital communication. Such systems would bridge the expressiveness divide by capturing and transmitting non-verbal signals in real-time, thereby enhancing the clarity and richness of virtual interactions.

Current methods for representing physical expressions in virtual environments often rely on manual input, such as selecting pre-defined animations or gestures. These methods limit expression representation to a small set of fundamental expressions such as anger, pleasure, fear, and sadness. While they are clearly significant, the human expression spectrum is considerably more diverse, spreading a broader variety of expressions encountered during social interactions. Although these approaches provide some level of expressiveness, they are limited by their static and user-dependent nature. Users must consciously choose how to represent their physical expressions, which can be cumbersome and may not accurately reflect their natural movements. In contrast, our research focuses on automating the process of physical expression detection and representation. By leveraging real-time facial expression

and upper body gesture recognition, the system dynamically translates users’ physical movements into avatar animations, ensuring that their expressions are conveyed accurately and effortlessly.

Collaboration in virtual environments holds immense potential for creating immersive and engaging social experiences [3]. However, a significant challenge remains: enabling users to express themselves physically in a natural and meaningful way. Existing research in this domain often focuses on either facial expressions or body language in isolation, neglecting the dynamic interplay between these elements that is fundamental to real-world interactions. Moreover, current systems typically limit physical expression to a narrow set of predefined gestures or animations. While these representations are useful, they fail to capture the full range of natural movements that occur during social interactions.

Despite advancements in avatar realism, many systems fall short of capturing the complexity of human physical expressions. Highly realistic avatars that replicate intricate facial details in real-time are still insufficient for conveying the depth and subtlety of physical communication. There is a need to move beyond static representations and incorporate the dynamic interplay of facial expressions, body language, and even subtle skeletal movements that underlie physical expressiveness. Furthermore, user-driven approaches, such as those used in platforms like Decentraland <sup>1</sup>, often rely on pre-defined animations that users manually select. While these methods provide some level of customization, they fall short of capturing the spontaneity and fluidity of natural physical interactions.

To address these limitations, our research focuses on the entire process of detecting and representing physical expressions in virtual environments. We came up with a pipeline that integrates real-time facial expression and upper body gesture recognition to enhance physical communication in virtual spaces. The system leverages machine learning algorithms to detect and classify users’ physical expressions based on their facial movements and body gestures. These detected expressions are then translated into dynamic avatar animations, enabling users to express themselves naturally and authentically. By automating the process of physical expression detection and representation, our system eliminates the need for manual input, ensuring that communication is seamless and accurate.

Our exploration of the entire process of capturing, detecting, and representing physical expressions in virtual environments. This includes introducing process by the design and development of a working model, the evaluation of its performance, and the analysis of its impact on user interactions. By focusing on the process, we aim to provide a comprehensive understanding of how physical expressions can be effectively integrated into virtual communication.

The introduced pipeline represents a significant step forward in bridging the expressiveness gap in virtual interactions. By capturing the dynamic interplay of facial expressions, body language, and skeletal movements, it provides a more comprehensive and nuanced representation of physical communication. This approach not only enhances the

---

<sup>1</sup><https://decentraland.org/>

clarity of virtual interactions but also fosters a greater sense of presence and connection among users. In an increasingly digital world, where virtual interactions are becoming the norm, our research offers a promising solution for improving physical expressiveness and collaboration in virtual environments.

### 2.1.1 Applications of Virtual Collaboration and Digital Socialization

Virtual collaboration and digital socialization have transformed how individuals and organizations interact, enabling immersive experiences that transcend physical boundaries. Below, we explore notable applications across different platforms, highlighting their impact on social, corporate, and entertainment sectors.

#### Party.Space: Corporate and Social Gatherings in the Metaverse

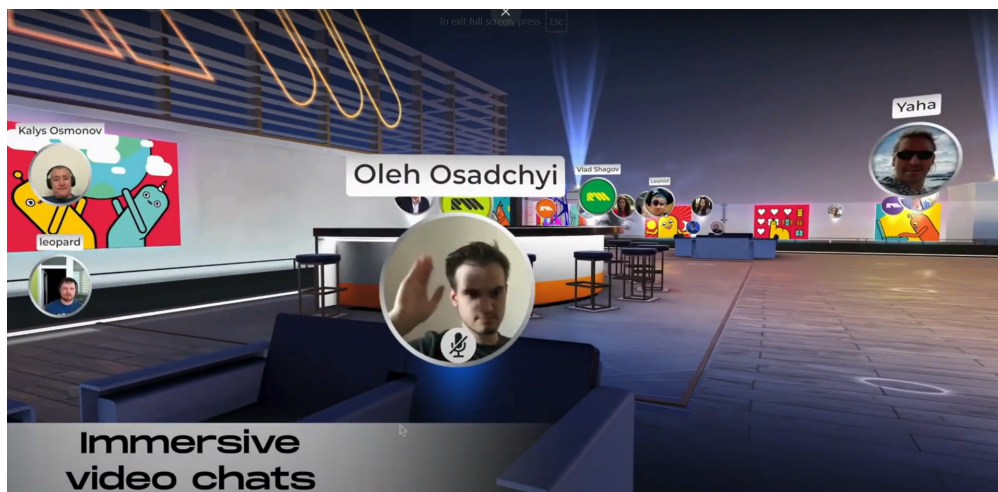


Figure 2.1: A Social gathering in Party Space where they use immersive video chats

As shown in the Figure 2.1 Party Space is a platform dedicated to hosting interactive virtual events, catering primarily to corporate celebrations, team-building activities, and entertainment. By leveraging the metaverse, it provides immersive experiences that bridge geographical gaps, allowing participants to engage in meaningful ways through avatars and virtual environments.

One of the key applications of Party.Space is facilitating metaverse parties for corporate milestones. For instance, Railsware, a software development company, celebrated its 15th anniversary with a virtual event featuring networking opportunities, gamified team challenges, and a digital awards ceremony. Similarly, Zapier, a fully remote company, hosted its first-ever employee retreat in the metaverse, enabling globally distributed teams to connect through avatars, participate in virtual escape rooms, and attend keynote speeches in a 3D space. Another example is SLOVA Tech PR's 8th-anniversary celebration, which included a virtual press conference, interactive booths highlighting company achievements, and an after party with live DJ performances.

Beyond corporate events, Party.Space also enhances sports and entertainment experiences in the metaverse. Juventus Academy Shanghai, for example, engaged fans by hosting a virtual training session where young athletes interacted with coaches via VR, took part in skill challenges, and accessed exclusive club content. Additionally, Party.Space has enabled synchronized movie watch parties, allowing attendees to stream films together, react in real-time using avatars, and discuss the content in virtual lounges. These innovative applications demonstrate how Party.Space is redefining social and professional gatherings in the digital age.

## **Decentraland: Large-Scale Virtual Events and Festivals**

Decentraland, a blockchain-based virtual world, has pioneered large-scale metaverse events, attracting global participation across industries.

Decentraland, a blockchain-based virtual world, has become a leader in hosting large-scale metaverse events, drawing participants from around the world across various industries. The platform has successfully created immersive experiences that blend culture, entertainment, and technology in a decentralized environment.

One notable example is Genesis City Art Week, a digital art exhibition that featured NFT galleries, live artist discussions, and interactive installations. Visitors could explore virtual galleries and purchase artwork directly within the metaverse. Another groundbreaking event was Metaverse Fashion Week (2023), where major brands like Dolce & Gabbana and Tommy Hilfiger presented digital wearables. Attendees could try on and buy virtual clothing for their avatars, merging fashion with blockchain technology in an innovative way.



Figure 2.2: A musical party in Decentraland

As shown in Figure 2.2 Music and entertainment have also thrived in Decentraland. The Decentraland Metaverse Music Festival (2022, 2023) included performances by renowned artists such as Björk and Ozzy Osbourne, set in elaborate virtual venues with interactive stages, VIP lounges, and NFT merchandise booths. Another unique experience was a stand-up comedy show where the audience's avatars could react in real-time with emojis, adding a dynamic layer to the performance.

Beyond entertainment, Decentraland has hosted events focused on community and advocacy. Metaverse Pride (2022) celebrated LGBTQ+ inclusivity with a virtual parade featuring themed floats, live performances, and educational discussions about diversity in Web3. Additionally, Wellness Week (2024) offered guided meditation, yoga classes, and mental health workshops in tranquil virtual settings, promoting well-being in the digital space.

The platform has also been a hub for gaming and tech innovation. The Decentraland Game Expo (2024) showcased upcoming blockchain-based games, allowing attendees to play demos, interact with developers, and earn NFT rewards. These events highlight how Decentraland continues to push the boundaries of what's possible in the metaverse, fostering creativity, connection, and commerce in a decentralized world.

### **Rec Room: Community-Driven Gaming and Social Interaction**

Rec Room is a platform that thrives on user-generated content, seamlessly blending gaming with social interactions to create engaging virtual experiences. One standout example is Geovanni's 17th birthday party, where friends from different countries gathered in a custom-built virtual space to play minigames as shown in Figure 2.3, exchange digital gifts, and celebrate together as if they were in the same room. Another popular activity is the Family Feud game night, which recreates the classic TV show in a multiplayer format, allowing families to compete in real-time quizzes with dynamic leaderboards, adding a fun and competitive twist to virtual gatherings.



Figure 2.3: A Social party where friends are planning to play pinball

Beyond social hangouts, Rec Room also serves as a space for game development and community-driven updates. Developers recently hosted a live playtest session for a new horror game demo, inviting users to explore a haunted mansion, share feedback, and earn exclusive in-game rewards. Another significant update was the release of Hotel HD, which introduced high-definition environments, upgraded voice chat features, and enhanced avatar customization, making social interactions more immersive and visually appealing.

These case studies highlight the expanding role of virtual collaboration in reshaping how people socialize, work, and entertain themselves. Hybrid events that merge physical and digital participation are becoming increasingly common, offering flexibility and broader accessibility. Corporations are also adopting these platforms more frequently to engage remote teams through interactive experiences. Additionally, the integration of NFTs into art, fashion, and gaming is growing, adding new layers of ownership and engagement in virtual spaces. As technology continues to advance, these platforms will evolve further, providing even more innovative and inclusive ways for people to connect, collaborate, and create shared experiences.

## **The Expanding Role of Virtual Collaboration**

The case studies presented reveal a fundamental shift in how people interact, work, and seek entertainment through virtual platforms. These digital environments are breaking down geographical barriers while creating new opportunities for connection that didn't exist before. One of the most significant developments has been the rise of hybrid events that seamlessly blend physical and digital participation. Companies now routinely host conferences where some attendees gather in person while others join via customizable avatars in meticulously designed virtual venues. Universities conduct graduation ceremonies where students walking across a physical stage see their digital counterparts celebrated simultaneously in a virtual replica of the campus. This hybrid approach maintains the energy of in-person gatherings while offering global accessibility, with features like real-time translation and AI-powered networking suggestions enhancing the experience for remote participants.

Corporate adoption of virtual platforms for remote team engagement has accelerated dramatically, particularly as companies recognize the limitations of traditional video conferencing. Forward-thinking organizations are moving beyond flat Zoom calls to immersive 3D workspaces where distributed teams can collaborate around virtual whiteboards, conduct product demonstrations in simulated environments, and even share casual coffee breaks in digital lounge areas. These platforms incorporate spatial audio that mimics real-world conversations, allowing natural side discussions to emerge during meetings. Team-building activities have been reimaged too - instead of awkward virtual happy hours, colleagues might compete in company-wide metaverse scavenger hunts or attend virtual cooking classes together, creating shared memories despite physical separation. The data shows these approaches lead to higher engagement, with some companies reporting 40% increases in participation for virtual events compared to traditional remote meetings.

Perhaps the most transformative trend has been the integration of NFTs and blockchain technology into virtual experiences, creating entirely new economic models for digital interaction. In the art world, virtual galleries now allow collectors to purchase NFT artwork that they can display in their personal metaverse spaces or even loan to public exhibitions. Fashion brands have embraced digital wearables, with limited-edition NFT outfits becoming status symbols that avatars can wear across multiple platforms. The gaming sector has been

particularly innovative, developing play-to-earn models where players truly own their in-game assets as NFTs that retain value outside a single game’s ecosystem. A concert attendee might purchase an NFT ticket that grants backstage access in the metaverse and doubles as a collectible, while a corporate training participant could receive an NFT certification verifiable on the blockchain. These developments are creating persistent digital identities and assets that travel with users across virtual spaces, fundamentally changing how we think about ownership and value in digital environments.

As underlying technologies continue advancing, we’re seeing the emergence of more sophisticated haptic feedback systems that allow users to “feel” virtual handshakes, AI-driven environments that adapt in real-time to user behavior, and lightweight VR hardware making immersion more comfortable for extended periods. The next frontier includes neural interface experiments that could one day translate thoughts into avatar expressions, and quantum computing applications that might enable massively complex virtual worlds with unprecedented realism. What began as simple virtual meeting spaces are evolving into persistent digital dimensions where work, social life, commerce, and creativity intersect in ways that are redefining human connection for the digital age. The organizations and individuals who master these new modes of interaction will gain significant advantages in engagement, collaboration, and cultural influence as this evolution continues.

## 2.2 Motivation

The rapid digitization of human interaction, accelerated by the COVID-19 pandemic, has fundamentally altered how we communicate, collaborate, and socialize. Virtual environments have become essential for work, education, and entertainment, yet they often lack the depth and expressiveness of face-to-face interactions. While text, voice, and video calls facilitate basic communication, they fail to fully replicate the nuanced non-verbal cues, facial expressions, gestures, and body language that are critical for meaningful human connection. Current virtual collaboration tools rely heavily on manual input for avatar expressions, limiting users to a narrow set of predefined expressions and gestures. This restriction stifles spontaneity and authenticity, making digital interactions feel artificial and detached. The challenge, then, is to develop systems that can automatically detect and translate real-world physical expressions into dynamic, real-time avatar animations, bridging the gap between physical and virtual communication. Our research is motivated by the need to enhance presence, engagement, and expression fidelity in virtual spaces. By leveraging machine learning for real-time facial and upper-body gesture recognition, we aim to introduce a pipeline that captures the full spectrum of human expressiveness, enabling avatars to reflect users’ natural movements and expressions without manual intervention. This advancement has far-reaching implications:

- **Improved Remote Collaboration:** More expressive avatars can enhance workplace communication, reducing misunderstandings and fostering stronger team cohesion in

distributed settings.

- **Enriched Social Experiences:** Virtual gatherings, from corporate events to gaming meetups, become more immersive when participants can convey expressions naturally.
- **Accessibility** Automated expression detection can benefit users with disabilities, providing alternative ways to engage in digital interactions.
- **Future-Proofing Digital Interaction:** As the metaverse evolves, realistic avatar expressiveness will be crucial for education, therapy, and even virtual commerce.

The applications discussed such as Decentraland’s music festivals, Party.Space’s corporate events, and Rec Room’s social gaming demonstrate the growing demand for richer virtual interactions. However, these platforms still rely on simplistic or manually controlled expressions, highlighting the need for more sophisticated solutions.

By advancing the automation of physical expression in virtual environments, our work seeks to make digital interactions as natural and expressive as real-life conversations. In doing so, we contribute to a future where virtual collaboration is not just a substitute for physical presence but an enhancement one that transcends geographical barriers while preserving the depth of human connection.



# Chapter 3

## Literature Review

### 3.1 Overview of Existing Work

The field of facial expression and gesture recognition has seen significant advancements in recent years, with numerous studies and technologies addressing various aspects of emotion detection and representation in virtual environments. [4] introduced the "Meta-EmoVis" research project, which uses the "Emoj Tool"<sup>1</sup> to map facial expressions to basic expressions and detect engagement levels as shown in Figure 3.1. However, this approach focuses solely on facial changes, neglecting the role of body gestures and postures in conveying expressions. Similarly, OpenFace 2.0, as highlighted by [5], provides a comprehensive framework for facial expression analysis using machine learning algorithms such as SVMs and CNNs. While OpenFace 2.0 is effective in capturing facial expressions, it does not account for body gestures, which are equally important in emotional communication.



Figure 3.1: Screenshot showing virtual emotional effects in the Meta-EmoVis project [4]

[6] explored the use of CNNs and FDCNNs to capture and represent expressions in the metaverse. Their work relies on datasets like the GEMEP Dataset and a recorded Emotion Dataset to train and validate models. However, their approach primarily focuses on body gestures and does not provide a mechanism for representing the detected expressions. Multiple studies utilize gestures to capture and express expressions, focusing on body

---

<sup>1</sup><https://emojlab.com/>

movements recorded by a Kinect Sensor. The work by [7] recognizes expressions including fear, happiness, sadness, calmness, and fury. However, this study only captures body movements and classifies them but does not represent the recognized expressions. The study by [8] implements a vector state representation by getting the points from the Kinect and computing the Euclidian distance between the vectors to identify the emotion. This method also focuses only on body gestures and doesn't have a representation mechanism. Moreover, executing complex movements with more than one body segment is challenging and it only considers discrete checkpoints. The research study by [9] proposes a system that supports a real human in communicating with a virtual human using natural body language and it considers 11 human upper body gestures with and without human-object interactions (be confident, have question, object, praise, stop, succeed and weakly agree). However, this method only considers static gestures and doesn't consider the facial expressions of the user. The study of [10] proposes an approach to real-time automatic emotion recognition from body movements and these are assessed on six emotion recognition problems (Happiness, Anger, Sadness, Surprise, Fear, and Disgust). However, this system only considers body gestures and needs other sophisticated equipment such as the Qualisys motion capture system. It is important to note that even though the Kinect Sensor, is an easily accessible commercial device that enables real-time emotion identification by recording and instantly processing body movements using classification algorithms, computing resources are needed for implementing these methods.

[11]'s study presents a systematic approach to understanding and measuring emotions in the growing metaverse. The study digs deeply into the gathering of physiological signals using non-invasive sensors, such as BVP, EEG, EKG, and GSR. This comprehensive method lays the groundwork for accurately representing the intricacies of emotional states experienced in virtual environments. By focusing on the choice of physiological variables, including GSR, BVP, EKG, and EEG, the research aims to identify patterns linked to various emotional states, enabling readings that are more accurate. A critical aspect of the research involves the development of a statistical multimodal model, integrating information from various biological signals, to robustly characterize basic emotional states and deepen the understanding of emotional dynamics in virtual environments. By focusing on three key emotional states along the arousal continuum- e-stress, e-engagement, and e-relax- the study elucidates underlying mechanisms and identifies unique physiological correlations for each condition. Additionally, the integration of real-time feedback mechanisms between tutors and students in online learning environments, based on emotional state categorization, is highlighted. Furthermore, the incorporation of participants' self-reported experiences with physiological data, utilizing standardized psychological measures like the STAI scale, enriches the comprehension of emotional states and streamlines validation processes. Finally, the utilization of the sympatho-vagal balance index, derived from EKG signals, allows for discerning minute differences between relaxation, stress, and engagement states, providing valuable insights into the functioning of the autonomic nervous system.

In [12]’s study, the utilization of biofeedback sensors in gaming explains their crucial role in indirectly capturing the emotional state of gamers, a concept applicable to quantifying emotional states within the metaverse emotion library. Within metaverse games, biofeedback sensors serve as indispensable tools for identifying physiological signals such as RESP, EMG, EKG, and GSR. These are highlighted in the study of [13] also. While these sensors don’t directly measure emotions, they indirectly monitor physiological changes associated with various emotional states. Through this indirect measurement, metaverse environments can derive quantitative information about players’ emotional reactions during gameplay. Parameters such as skin conductance or heart rate variations offer insights into emotional responsiveness and arousal levels, facilitating the measurement of players’ emotional states. Despite their indirect nature, biofeedback sensors provide valuable insights into physiological arousal and reactivity, closely linked with emotional states. Through advanced analysis of these physiological data, researchers can develop customized algorithms to deduce and measure affective reactions, contributing to the development of a comprehensive emotional repository within the metaverse. This approach underscores the importance of integrating advanced technology to enhance user experiences and emotional engagement within virtual environments.

Moreover, according to [14], AI systems in AR use techniques like facial expressions and physiological signals (EEG and ECG) to convey expressions. Deep neural networks analyze these inputs to identify and simulate expressions. Facial emotion systems use cameras for image processing, while systems like AuRea use sensors for ECG data. AR systems can detect expressions such as fear, surprise, anger, sadness, and happiness. Facial recognition systems are more accessible, whereas physiological signal-based systems require advanced setups. Also, these systems tend to capture expressions rather than represent them.

[15] also suggest a method to detect expressions from gestures using skeletal data obtained from Kinect-like devices as input and textual description of their meaning. This method uses the DTW method for gesture recognition and by following dynamic algorithms it maps the real-time gesture with previously recognized gestures and via that identifies the relevant emotion that correlates to the user’s spoken sentences. However, this method mainly focuses on body gestures and spoken sentences by users and does not focus much on facial expressions.

In addition to the above work, [16] proposed a method for emotion classification from a comprehensive list of body movement features. The classification step employs a two-layer feature selection framework. The first layer combines ANOVA and MANOVA to discard irrelevant features. In the second layer, a binary chromosome-based genetic algorithm is used to choose a feature subset from the relevant features, enhancing the emotion recognition rate. This method concentrates solely on body gestures, capturing user movements to identify expressions without representing the recognized expressions.

The work by [17] suggests a technique to recognize expressions automatically by body gesture analysis. Here, it mainly focuses on the motion of the upper body (head and hands).

This work has developed a model and algorithms to analyse expressive content and it has individuated which motion cues are involved in conveying actors' expressive intentions to represent four expressions; anger, joy, relief, and sadness through a scenario approach. However, this method extracts movements from the GEMEP dataset as shown in Figure 3.2 and it does not extract body movements from users interacting in a virtual collaborative context.

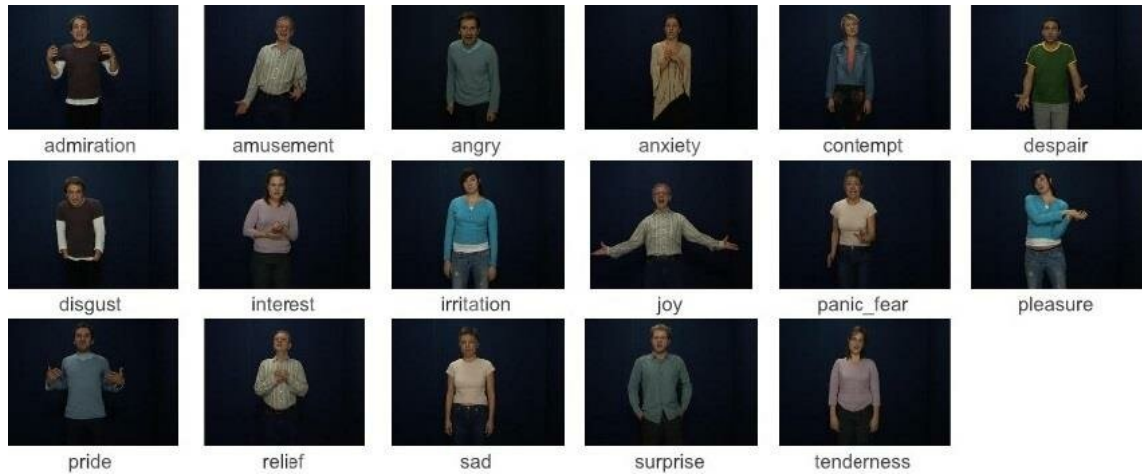


Figure 3.2: Screenshot of GEMEP Data Set [18]

[19] suggests a novel solution to automatically recognize facial expressions in the context of group meetings. This study mainly focuses on users' smiles and it aims to recognize who is smiling at whom, when they are smiling, and how often they smile. The authors propose a novel algorithm that automatically selects the interest points and jointly estimates facial pose and expression in the framework of the particle filter. Even though this approach captures expressions in a collaborative environment it does not represent any emotion back and also it only focuses on facial expressions.

[20] propose a solution that learns expressions through upper body movements and corresponds with facial expressions. It follows two approaches: a CNN to classify expressions without considering temporal features and an LSTM to classify expressions based on temporal information. This method uses the FABO dataset shown in Figure 3.3 and the FER-2013 dataset as benchmarks. Even though this method considers both facial expressions and body gestures and doesn't require sophisticated equipment, it doesn't focus on emotion expression representation.

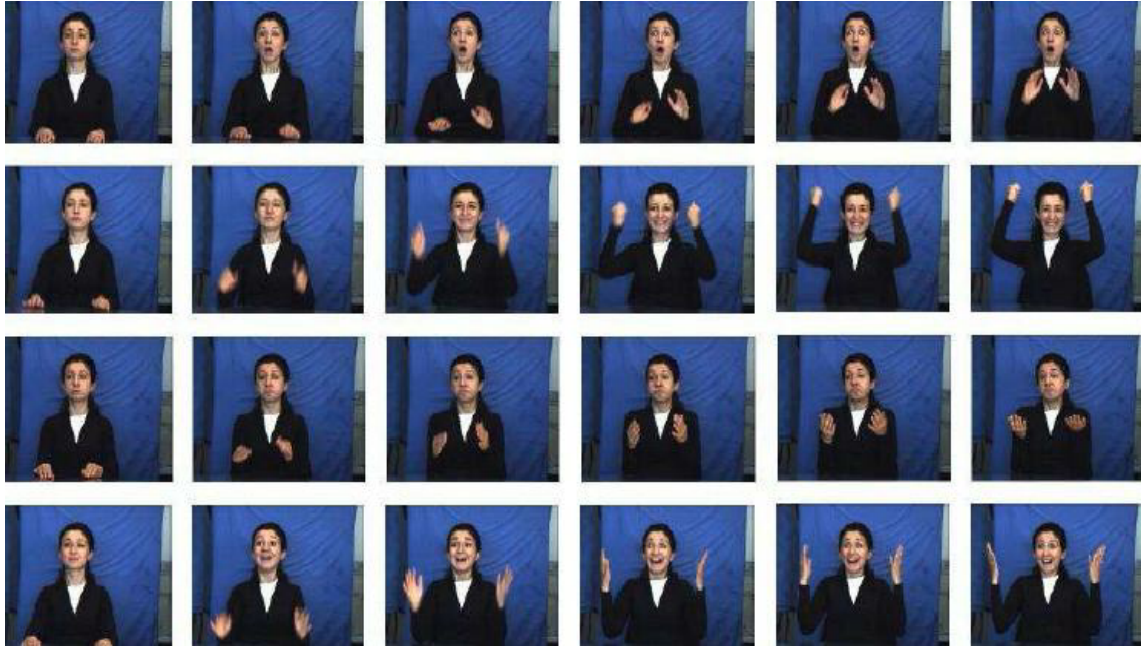


Figure 3.3: Example gestures from the FABO Data Set [21]

Additionally, the study by [22] presents an approach to automatic visual recognition of expressive face and upper-body gestures from video sequences suitable for use in a vision-based affective multi-modal framework. It uses two cameras, one to capture facial expressions (anger, disgust, fear, happiness, sadness, and surprise are the six basic prototypical facial expressions) and the other to capture body gestures (static body posture related to anger, disgust, fear, happiness, sadness, and surprise). This method only considers static body postures and there is no enhanced representation mechanism.

When focusing on expressing user emotions and feedback within the metaverse, it is important to detect them accurately and efficiently to increase realism. Using VR headsets alone is insufficient for correctly capturing user emotions and body gestures. As a result, researchers are using a variety of infrastructures to address this difficulty, each with its own method of identifying and interpreting human attitudes within immersive digital worlds.

One such approach is the incorporation of OpenCV, a potent computer vision library, within the metaverse’s architecture. [23] states that with OpenCV’s extensive toolkit for image analysis and processing, OpenCV makes it possible to follow and identify user movements, gestures, and facial expressions in real time. This enhances the emotional depth and interactive nature of avatar-based interactions in virtual environments by enabling the development of immersive experiences that dynamically react to user actions. The study also examines current technologies and offers an analysis of their benefits and drawbacks, including game consoles with gesture controllers and specialist cameras like Intel RealSense. The study shows that OpenCV’s incorporation into the metaverse marks a substantial development in avatar interaction paradigms, with the potential to completely transform the way users interact with virtual environments by precisely recognizing and interpreting their emotions and body language.

One of the biggest challenges in the Metaverse is precisely capturing emotions and

bodily gestures in order to achieve real emotional expression. In order to tackle this issue, researchers and programmers are investigating a range of innovative technology-based infrastructures, with gaming engines such as Unity and Unreal Engine being major participants. According to [24], these engines provide a full range of capabilities and tools necessary to build realistic virtual worlds in which avatars may communicate and effectively portray emotions. With Unity and Unreal Engine’s advanced character animation and rigging features, developers can give avatars life and have them display a wide variety of emotions through realistic gestures and movements. Furthermore, these engines’ support for physics simulation guarantees that avatar interactions with the virtual world feel authentic and natural, which heightens the immersive experience. Most importantly, Unity and Unreal Engine’s user-generated content and customization options let developers incorporate sophisticated interaction mechanics like gesture recognition and facial expression tracking, which makes it easier to accurately record body language and emotions in the Metaverse. Furthermore, these engines’ cross-platform interoperability allows avatars and users to interact with each other seamlessly across various virtual settings, guaranteeing a consistent and immersive experience across platforms.

### **3.1.1 Literature on Facial and Body Movements Related to Expressions**

The table 3.1 summarizes key findings from various studies on gestures and facial expressions in communication. These papers explore how body language, such as hand movements, posture, and facial cues, can convey expressions, intentions, and expressions. The insights highlight the nuanced ways non-verbal behaviour influences interpersonal interactions.

<b>Paper</b>	<b>Information on gestures</b>	<b>Information on facial expressions</b>
The Influence of Non-Verbal Behaviour on Meeting Effectiveness and Pro-Active Behaviour: A Video Observational Study (2017) [2]	<p>Having the palms downwards means that the person wants to interrupt a conversation or stop it due to disagreement or irrelevance.</p> <p>Palm down gestures are used in contexts of denial or interruption.</p> <p>Palms upwards imply offering or requesting.</p> <p>Hands moving apart with palms up imply withdrawal.</p> <p>Self-adaptors (touching oneself) show worry or fear.</p> <p>Touching hair implies worry.</p>	<p>People that are feared hide their faces in both hands or clasp them together.</p> <p>No further indication.</p>
“Actions Speak Louder Than Words” - Body Language in Business Communication (2009) [25]	<p>Leaning back and closed - lack of interest.</p> <p>Leaning back and open - contemplation.</p> <p>Leaning forward and closed - potential aggression.</p> <p>Leaning forward and open - interest/agreement.</p> <p>Head neutral - open attitude.</p> <p>Tilted back - superiority.</p> <p>Tilted down - judgment.</p> <p>Tilted to side - interest.</p>	<p>No indication.</p>

Paper	Information on gestures	Information on facial expressions
Bi-modal emotion recognition from expressive face and body gestures (2006) [22]	<p><b>Anxiety:</b> Hands near table; fingers move/tap.</p> <p><b>Anger:</b> Body extended; hands on waist/fists low.</p> <p><b>Disgust:</b> Body backing; hand on neck/face.</p> <p><b>Fear:</b> Contracted body; hands raised to cover.</p> <p><b>Happiness:</b> Extended body; fists high.</p> <p><b>Uncertainty:</b> Shoulder shrug; palms up.</p>	<p><b>Anxiety:</b> Lip bite, mouth stretching, eye movement, lip wipe.</p> <p><b>Anger:</b> Brows lowered, tense lids, firm lips.</p> <p><b>Disgust:</b> Upper/lower lip movement, wrinkled nose, tongue out.</p> <p><b>Fear:</b> Brows raised/drawn, wrinkled forehead, mouth open.</p> <p><b>Happiness:</b> Lip corners up, cheek/lid wrinkles.</p> <p><b>Uncertainty:</b> Brow movement, chin/jaw shifts, lip corners down.</p>
Preliminary analysis of facial expressions and body movements of four types of laughter (2024) [26]	<p><b>Mirthful/Boosting Laugh:</b> Body leans forward then returns; rhythmic shoulder trembling; head moves downward/forward.</p> <p><b>Smoothing/Softening Laugh:</b> Still upper body; fewer head moves; more single shoulder movements.</p>	<p><b>Mirthful/Boosting Laugh:</b> Raised cheeks; closed/narrow eyes; eye corner wrinkles; wide open mouth.</p> <p><b>Smoothing/Softening Laugh:</b> Slightly raised cheeks; normal eye openness; fewer wrinkles; mostly closed mouth.</p>

Table 3.1: Facial Expressions and Body Gestures Identified from Literature for Different Expressions

## 3.2 Theoretical Framework

The theoretical framework of this research is grounded in the integration of facial expression and gesture recognition to enhance physical expressions in virtual environments. The study builds on the concept of multimodal expression recognition, which combines multiple sources of information, such as facial expressions and body gestures, to improve the accuracy and robustness of emotion detection. This approach is supported by the work of [20], who used CNNs and LSTMs to classify expressions based on both facial expressions and body gestures.

Additionally, the research draws on the principles of HCI and affective computing, which emphasize the importance of natural and intuitive interfaces for enhancing user experience.



By automating the detection and representation of physical expressions, the proposed system aligns with these principles, providing a seamless and immersive virtual communication experience.

### 3.3 Critical Analysis

The primary critique directed at the metaverse area is that it has a propensity to place technology innovation above a thorough comprehension and moral consideration of human experiences, especially expressions. Although capturing and representing expressions is the goal of many metaverse systems, these approaches frequently operate within a limited technical framework, ignoring the richness and contextuality of human emotional expression. Furthermore, the focus on immersive experiences and real-time engagement may overshadow the significance of consent and privacy in the recording and portrayal of expressions. Furthermore, the accessibility and inclusiveness of emotion identification systems may be limited by the need for complex laboratory equipment and specialized tools, which might put researchers and developers at a disadvantage.

With its ability to capture a broad range of expressions, such as surprise, happiness, sorrow, anger, contempt, fear, and neutral states, Meta-EmoVis provides a thorough method for emotion detection. Emojis and a colourful halo over the avatar offer a visually appealing depiction. However, by ignoring other possible signs like physiological signals or body motions, its emphasis on facial expressions restricts the range of emotion capture. Moreover, the absence of a method for storing collected expressions for later analysis in Meta-EmoVis may limit its usefulness in long-term studies or applications that need historical data.

The Framework for recording and synchronizing experiences in virtual reality experiences with physiological signals offers a viable path for capturing expressions through the integration of many physiological signals, including heart rate, body acceleration, EDA, EMG, and breathing rate. This method provides a comprehensive understanding of emotional states, but it ignores the direct expression of certain expressions due to its major focus on physiological data. Furthermore, its applicability and accessibility in real-world scenarios may be limited by the need for complex laboratory infrastructure.

OpenFace 2.0 distinguishes itself with its real-time emotion identification from head and facial positions and its simplicity of accessibility. It offers insightful information on the user's emotional state by identifying a variety of expressions. However because the technology only looks at facial expressions, it ignores other possible clues like body language or physiological cues. Furthermore, OpenFace 2.0 does not have an emotion storing feature, which limits its use in longitudinal research or applications that need to analyze previous data.

The study on body gestures and emotion identification using kinetic sensors demonstrates how expressions may be captured through upper-body movements. This technique is new, but it can only be applied to a restricted range of emotional states and circumstances due to its reliance on kinetic sensors and concentration on a small number of expressions. Its usefulness for long-term research or applications needing historical data is further limited by

the lack of emotion storing features.

Kinect motion capture data of human gaits used for emotion detection shows a multimodal approach by recognizing expressions from both physiological signs and facial expressions. Practical restrictions are presented by the lack of real-time analytical capabilities and the dependence on expensive laboratory infrastructure. Its inability to store expressions also makes it less useful for long-term studies or applications that need previous data.

A promising avenue for identifying expressions from both facial and body gestures is the deep learning approach to emotion identification from human body movements. Though it has potential, its practical use may be hampered by the need for complex laboratory infrastructure and the lack of real-time analysis capabilities. Additionally, its usefulness for longitudinal research or applications needing previous data is limited by the absence of emotion storing capabilities.

The goal of "Automatic Quantitative Evaluation of Emotions in E-learning Applications" is to employ non-invasive sensors to measure emotions in e-learning environments. Although this method provides a methodical way to evaluate emotions, subtleties in emotional states may be missed due to its emphasis on certain metrics like e-stress, e-engagement, and e-relaxation. Moreover, its lack of emotion storage limits its use in longitudinal research or applications that need to analyze past data.

The integration of physiological arousal and reactivity for gaming interaction is explored in the study "Using Direct and Indirect Physiological Control to Enhance Game Interaction". Although this method provides a fresh viewpoint on interactions motivated by emotion, it ignores other possible cues such as body language or facial expressions by concentrating only on physiological reactions. Moreover, its usefulness for long-term research or applications needing historical data is limited by the lack of emotion storage capabilities.

In conclusion, there is still a strong need for a more nuanced and ethically aware approach, even while developments in emotion capture and representation inside the metaverse show great potential for improving virtual experiences and interactions. In order to create more inclusive, accessible, and ethical technologies, researchers and developers must first recognize the shortcomings and difficulties present in current techniques.

### **3.3.1 Identifying Gaps in the Literature**

Despite the vast literature on technical improvements and societal upheavals in the metaverse, numerous crucial gaps remain, notably in the thorough detection and representation of expressions.

Firstly, while several studies investigate technology breakthroughs such as holographic displays, micro-LEDs, and 5G connectivity infrastructure, there is a significant lack of study on the integration of multiple modalities for emotion identification inside virtual worlds. Existing methods typically focus on facial expressions or physiological data, ignoring the possible synergies and increased accuracy that may be gained by merging both modalities.

A comprehensive strategy that includes face and bodily expressions, as well as physiological indicators like heart rate and electrodermal activity, is required to capture the depth and complexity of human expressions in the metaverse.

Furthermore, there is a significant gap in research in the area of storing and retrieving emotional data in a central library or database. While some studies record expressions in real time, they frequently lack the means for storing and interpreting this information beyond its immediate utility. A unified library of expressions would make it easier to do long-term studies, identify trends, and personalize metaverse experiences. By keeping emotional data in an organized and accessible way, researchers and developers may get new insights into user behavior and preferences, resulting in more immersive and engaging virtual experiences.

Furthermore, the literature lacks research into approaches for augmenting and enhancing expressions in virtual worlds. While some research focuses on emotion identification, few look into ways for dynamically changing users' emotional states in real-time. Techniques like regulating expressions, biofeedback, and personalised information distribution have the potential to improve user experiences and promote emotional well-being in the metaverse. Future studies should investigate these strategies' effects on user engagement, contentment, and overall immersion.

Finally, there has been little study on the real-time representation and feedback of expressions in virtual worlds. While some research collects expressions in real time, few offer participants rapid feedback or visualizations of these expressions. Real-time emotion representation via avatars, animations, or other visual signals can improve social interaction, communication, and empathy in the metaverse. Virtual environments that provide users with real-time feedback on their emotional states can increase self-awareness, emotional control, and interpersonal connection.

Related Work	Are expressions captured or represented?	What is represented	What is captured? Facial/Body movements	What expressions are detected	Enhancement of expressions
Meta-EmoVis	Captured and represented	Surprise, happiness, sadness, anger, disgust, fear and neutral	Facial	Surprise, happiness, sadness, anger, disgust, fear and neutral	The halo on top of the avatar changes its colour, Emojis spread from avatar's body
Framework for recording and synchronizing experiences in VR with physiological signals	Captured	N/A	Body	Heart rate, body acceleration, electrodermal activity (EDA), electromyography (EMG)	No
OpenFace 2.0	Captured	N/A	Facial/head poses	Anger, disgust, fear, happiness, sadness, surprise, contempt, valence, arousal	No
A study on Emotion recognition from body gestures using kinetic sensor	Captured	N/A	Upper body	Anger, fear, happiness, sadness and relaxation	No
Emotion recognition using Kinect motion capture data of human gaits	Captured	N/A	Facial expressions/ Physiological signals	happiness, sadness, anger, surprise, and fear	No
Deep learning approach for emotional recognition from human body Movements	Captured	N/A	Facial/Body	happiness, anger, sadness, fear	No
Automatic Quantitative Evaluation of Expressions in E-learning Applications	Captured	N/A	Body	e-stress, e-engagement, e-relax	No
Using Direct and Indirect Physiological Control to Enhance Game Interaction	Captured	N/A	Body	physiological arousal and reactivity	No

Table 3.2: Characteristics related to expressions and their analysis

Related Work	Infrastructure used to Represent	Infrastructure used to capture	Easily Accessible or Sophisticated Laboratory Infrastructure	Real-time or not	Whether expressions are stored or not in a library	Whether the library is globally available or not
Meta-EmoVis	Coloured halo on top of the avatar, Camera, Emojis	Emoj tool	Easily Accessible	Real-time	No	No
Framework for recording and synchronizing experiences in VR with physiological signals	N/A	BiosignalsPlux Hybrid-8 module	Sophisticated Laboratory Infrastructure	Real-time and offline	No	No
OpenFace 2.0	N/A	OpenFace 2.0 Toolkit	Easily Accessible	Real-time	Yes	No
A study on Emotion recognition from body gestures using kinetic sensor	N/A	Kinetic sensor	Easily Accessible	Real-time	No	No
Emotion recognition using Kinect motion capture data of human gaits	N/A	Camera, devices to capture ECG/ EEG data	Sophisticated Laboratory Infrastructure	No	No	No
Deep learning approach for emotional recognition from human body Movements	N/A	Camera - Images / videos	Sophisticated Laboratory Infrastructure	No	No	No
Automatic Quantitative Evaluation of expressions in E-learning Applications	N/A	Non-invasive sensors	Sophisticated Laboratory Infrastructure	Real-time	No	No
Using Direct and Indirect Physiological Control to Enhance Game Interaction	N/A	Sensors	Sophisticated Laboratory Infrastructure	Real-time	No	No

Table 3.3: Characteristics related to infrastructure, accessibility, and storage

To summarize, while the metaverse literature provides essential insights into technical breakthroughs and societal transitions, significant gaps exist in the thorough identification and representation of expressions. Future research should concentrate on integrating several modalities for emotion identification, creating centralized libraries of emotional data, investigating approaches for emotion enhancement, and providing real-time representation and feedback of expressions within virtual settings. Addressing these gaps will not only help us better comprehend human expressions in the metaverse, but it will also improve the overall user experience and promote emotional well-being. Table 3.2 shows a comparison of related work according to expressions and their analysis and Table 3.3 shows a comparison of infrastructure, accessibility and storage.

### 3.4 Relevance to the Research

This research builds on the existing body of work in facial expression and gesture recognition by focus to physical expressions. While many studies, such as those by [4] and [5], concentrate on detecting or representing expressions, our work emphasizes the accurate capture and representation of physical expressions, including facial movements and upper body gestures. This approach ensures that the system reflects the user’s natural movements and interactions, rather than inferring emotional states, which can be subjective and context-dependent.

The this research focuses on the entire process of capturing, detecting, and representing physical expressions in virtual environments. Unlike previous studies that rely on specialized equipment, such as the Qualisys motion capture system [10] or multiple cameras [22], our

system leverages widely accessible technologies, such as webcams and machine learning algorithms. This makes the system practical and scalable for real-world applications, addressing a significant limitation in existing research.

The process of this research involves several key steps:

1. Understanding Collaborations in Virtual Worlds
2. Recognition of Expressions Expressed During Collaboration
3. Detection and Classification of Facial Expressions and Body Gestures
4. Creation and Exaggeration of Expressions Based on Skeletal Point Patterns and Facial Landmarks in form of animations
5. Creating a near Real-Time Facial and Body Gesture Input mechanism
6. Identifying Skeletal Point Sequences and Patterns in near real time
7. Mapping with Predefined expression animations
8. Expressing Exaggerated expressions through Avatars in the virtual world
9. Assessment of How the Automated Gesture-Based Solution Enhances User Engagement and Communication

# Chapter 4

## Methodology

This research adopts the **DSRM**, a structured framework for developing and evaluating practical solutions to real-world problems. Unlike traditional observational research, DSRM emphasizes the creation of artifacts, in this case, a real-time facial expression and gesture detection system integrated with a virtual avatar environment. The methodology is problem-driven, iterative, and evaluation-focused as shown in Figure 4.1, ensuring that the developed solution is both technically robust and practically valuable.

The DSRM process typically involves the following iterative steps, which guided this research.

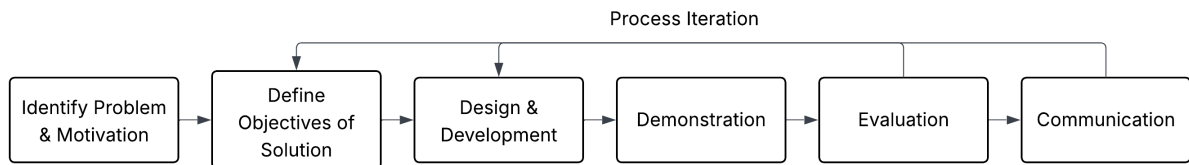


Figure 4.1: DSRM Steps [1]

1. **Problem Identification and Motivation:** This initial step involves identifying a relevant and important problem in a real-world setting that can be addressed through the design and development of an artifact. It also includes motivating the research by explaining the significance of the problem and the value of finding a solution.

2. **Definition of Objectives for a Solution:** Based on the identified problem, this phase defines the objectives that the designed artifact should achieve. These objectives should be clear, specific, and measurable, outlining what the solution is intended to do.

3. **Design and Development:** This is the core of DSRM, where the artifact is actually created. This can involve developing constructs, models, methods, or instantiations (software systems, algorithms, etc.). This phase is often iterative, involving building and refining the artifact based on feedback and evaluation.

4. **Demonstration:** In this step, the developed artifact is used to solve or address the identified problem. This demonstrates the artifact’s functionality and its potential to provide a solution in a relevant context.

5. **Evaluation:** This phase involves assessing how well the artifact achieves the objectives defined earlier and how effectively it solves the problem. Evaluation can employ various methods, both quantitative and qualitative, to measure the artifact’s performance and impact. This step often leads to further iterations of the design and development process.

6. **Communication:** The final step involves communicating the research, including the problem, the artifact, its demonstration, and its evaluation, to the relevant audience (e.g., researchers, practitioners) through publications, presentations, or other means. This contributes to the knowledge base and allows others to build upon the research.

These steps are often followed in an iterative manner, meaning that researchers may cycle back to earlier steps based on the findings from later steps, particularly during design and evaluation. The emphasis is on creating a useful artifact and rigorously evaluating its effectiveness in addressing a real-world problem.

## 4.1 Why DSRM?

The DSRM provides a structured approach to develop and evaluate a novel process for enhancing physical expressiveness in virtual collaborative environments. Unlike commercial product development, our focus centers on establishing and validating a comprehensive pipeline that bridges the gap between real-world physical expressions and their digital representations. This methodological choice is particularly apt because it allows us to systematically address the complex challenge of capturing, interpreting, and translating human expressions into virtual environments through a series of carefully designed and evaluated process stages.

The research begins with identifying critical use cases that reveal the limitations of current virtual expression systems. In virtual meetings, for instance, participants frequently struggle to convey nuanced feedback or expressive responses through static avatars, often resulting in miscommunication during sensitive discussions like performance reviews or creative brainstorming sessions. Social Metaverse-based platforms present another compelling use case, where the lack of spontaneous physical expressiveness inhibits the formation of meaningful personal connections, making interactions feel artificial and constrained. These use cases collectively highlight the pressing need for a more sophisticated approach to physical expression in virtual spaces.

The choice of DSRM stems from its problem-centered nature, which perfectly matches our starting point of identifying significant deficiencies in current virtual platforms. Where standard communication tools and metaverse environments rely on manual input or simplistic avatar animations, resulting in constrained and unnatural interactions, DSRM guides researchers through a rigorous process of developing practical solutions to such recognized issues. The methodology’s emphasis on creating functional artifacts makes it ideal for our work, as we need to produce an actual system that integrates facial expression recognition, upper body gesture detection, and dynamic avatar animation - components that must work together seamlessly to achieve more authentic virtual communication.



The methodology then informs our systematic process design phase, where we establish the framework for real-time expression capture and translation. Rather than building a commercial product, we focus on creating a modular process architecture that specifies how facial recognition algorithms and models interface with gesture detection systems, how these inputs are processed and filtered, and how they ultimately drive avatar animations. This process-oriented approach allows for various implementation possibilities while maintaining core principles of expression fidelity and system responsiveness. DSRM’s emphasis on artifact design, in our case, translates to designing this comprehensive process flow, which is complete with decision points and quality control mechanisms at each stage.

For the development phase, DSRM guides our creation of process prototypes – not as market-ready solutions, but as demonstrators of the proposed methodology. We develop limited-scale implementations to validate each process component: a facial expression and upper body gesture pipeline that tracks key skeletal movement sequences and an animation system that synthesizes these inputs into coherent avatar expressions. Each prototype serves to test and refine our pipeline specifications, ensuring they can accommodate the variability observed in our use cases. The corporate meeting scenario, for example, helps validate the process’s ability to distinguish between similar expressions, while the social VR context tests its capacity for rapid expression transitions.

Evaluation under DSRM occurs at multiple process levels. Technical validation assesses whether each process stage meets its design specifications in terms of accuracy and performance. For instance, can the expression detection process reliably identify expressions of the user from a range of dynamic facial expressions and upper body gestures? Can the animation process maintain natural-looking movements when handling rapid expression changes during lively virtual social gatherings? Can the framework built be used without any sophisticated infrastructure? User evaluations then examine whether the end-to-end process actually enhances communication effectiveness in our target use cases by evaluating whether the built process increases the user’s collaborative experience. These evaluations provide critical feedback for process refinement.

The iterative nature of DSRM proves particularly valuable for our process-focused research. Each cycle allows us to adjust pipeline parameters based on findings from both technical and user evaluations. We might modify the expression detection thresholds after observing corporate users’ needs for subtler professional cues, or adjust the animation smoothing algorithms based on social VR participants’ preferences for more pronounced expressions. This iterative refinement continues until the pipeline demonstrates consistent effectiveness across our primary use cases while maintaining technical feasibility.

DSRM’s knowledge contribution aspect aligns perfectly with our goal of advancing understanding in virtual communication processes. The methodology helps us articulate generalizable principles for physical expression translation in virtual environments, such as optimal expression sampling rates, effective mapping strategies between real-world expressions and avatar representations, and trade-off considerations between expression

fidelity and system performance. These contributions extend beyond any single implementation, providing valuable guidance for future research and development in this field.

The methodology also accommodates our research’s exploration of boundary conditions – understanding where and when our proposed pipeline succeeds or faces limitations. Through systematic testing across our use cases, we can identify, for example, whether the pipeline works equally well for different types of expression representations. These insights form an important part of our process knowledge contribution, helping define the scope of applicability for our approach.

Ultimately, employing DSRM allows us to develop and validate a robust process framework that addresses the expressiveness gap in virtual collaboration while acknowledging the research’s exploratory nature. The methodology provides the necessary structure to ensure our pipeline design is rigorous and well-grounded, while its flexibility accommodates the uncertainties inherent in developing novel approaches to complex communication challenges. By focusing on the pipeline rather than a commercial product, we create knowledge that can inform various future implementations while advancing academic understanding of virtual communication dynamics.

## **4.2 Main Activities Carried Out**

### **4.2.1 Problem Identification and Motivation**

The initial phase of this research, problem identification and motivation, involved a systematic investigation into the limitations prevalent in current virtual communication platforms. To achieve this, a comprehensive literature review was conducted, examining studies published between 2008 to 2023. The review primarily focused on research concerning systems designed for expression identification through facial expressions and body gestures, as well as those aimed at enhancing realism in virtual environments such as the Metaverse.

In parallel with the literature review, an exploratory analysis of existing applications was performed. This included examining platforms like Decentraland to understand their approaches to expression representation. Additionally, widely used communication applications such as WhatsApp in iOS devices as shown in Figure 4.2 and Microsoft Teams were investigated to identify how they incorporated automatic gesture recognition in video call functionalities.

A critical analysis framework was developed to evaluate the capabilities of existing systems and applications. This considered five key factors: (1) whether expressions were captured and/or represented; (2) what is represented in the virtual environments; (3) whether facial movements, body movements, or both were captured; (4) what expressions are detected; and (5) whether any enhancements to expressions were implemented.

The critical analysis revealed a significant gap in existing solutions: no single system comprehensively addressed all five of these critical factors. Many systems either focused



Figure 4.2: Static Gesture Recognition in WhatsApp in iOS devices

solely on capturing expressions without robust representation mechanisms or relied on manual input for repress expressions through avatars. Furthermore, the analysis indicated that most existing solutions for gesture recognition were limited to identifying static gestures, such as a "thumbs up" or a "victory" sign. In many cases, users were required to manually select an icon to trigger an avatar's gesture representation. Another notable limitation identified was the dependence of some solutions on sophisticated and often expensive infrastructure, such as Kinect devices, which limits their accessibility for a broad user base.

Considering these identified limitations and the clear need for more natural and intuitive virtual interactions, the motivation for this research emerged. The primary motivation was to develop a pipeline that would comprehensively address the shortcomings identified in the critical analysis. Specifically, we were driven to build a system capable of capturing and interpreting the full range of movements naturally expressed by a user seated in front of a computer during a virtual collaboration session. A key aspect of our motivation was to create a solution that would be easily accessible and not require any sophisticated or specialized hardware, allowing any user with a standard webcam or built-in camera to utilize the system effortlessly. This problem identification and subsequent motivation laid the groundwork for defining the research objectives and the subsequent design and development of the proposed solution.

With that in mind, three research questions were planned to address,

1. What specific facial expressions and body gestures convey distinct expressions, and what are the corresponding sequences of skeletal points associated with these gestures?
2. In a real-time skeletal point sequence, how can we identify predefined facial expressions and body gesture patterns within near real-time and express them?
3. To what extent will the suggested solution perform and help users facilitate collaborative interactions in the virtual space?

## **4.2.2 Iteration 1 (May 2024 to November 2024)**

### **Definition of Objectives for a Solution**

To address the above-mentioned limitations, a pipeline was planned that would capture users' facial expressions and upper-body movements and represent them as exaggerated expressions through an avatar in a virtual environment within near real-time. The seven objectives mentioned below was drafted to be satisfied by this study.

1. Understanding which expressions are expressed and identified in a collaborative environment.
2. Understanding the connection between facial expressions, body gestures, and skeletal point sequences in conveying expressions.
3. Capturing and analyzing data on skeletal point sequences in conjunction with facial expressions in displaying expressions.
4. Understanding on how a virtual avatar expresses above-identified expressions.
5. Building and evaluating an automated system that uses these patterns for expression recognition and presentation in a virtual environment
6. Design a method to represent expressions with varying intensities on an avatar in the virtual space.
7. Evaluate the effectiveness and performance of the system in facilitating user collaboration.

### **Design and Development**

#### **System Design:**

The design and development phase of this research focused on creating a robust and accessible system capable of accurately capturing, interpreting, and representing physical expressions in a virtual environment. The system design commenced with the decomposition of the overall functionality into three key components as shown in Figure 4.3.

1. Component One: Responsible for capturing facial expressions and upper-body gestures and identifying the expression

2. Component Two: Responsible for representing the identified expression as an expression via an avatar
3. Component Three: Responsible for integrating components 1 and 2.

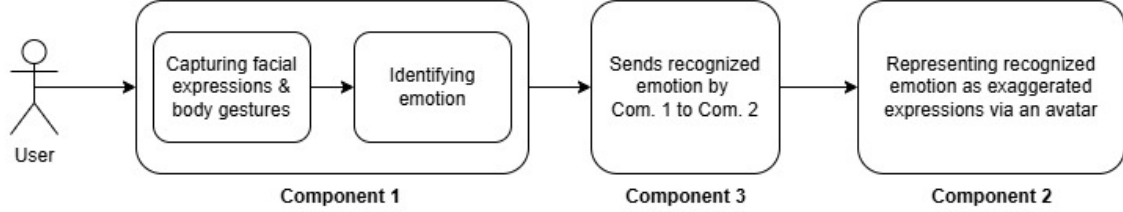


Figure 4.3: Key Components of VirExp

Following the identification of these core components, a thorough investigation was conducted to determine the most suitable technologies for their implementation. This pipeline involved reviewing existing literature, exploring open-source frameworks and libraries, and considering their effectiveness and accessibility.

### Identifying Potential Technologies:

The study [20], which used a CNN to extract facial and body movements, served as our starting point for the first component. Therefore, we carried out research for such CNNs already available and found the MediaPipe Holistic Library <sup>1</sup> provided by Google. The MediaPipe Holistic library provides a comprehensive solution for capturing skeletal points by combining pose, face, and hand landmarks into a unified human body model as shown in Figure 4.4. This library is ideal for analyzing full-body gestures, poses, and actions using an ML model applied to a continuous stream of images, making it highly effective for real-time applications. It produces 543 landmarks in total, covering 33 pose landmarks, 468 face landmarks, and 21 hand landmarks for each hand. Since VirExp only considers facial expressions and upper body gestures, MediaPipe Holistic is well-suited, as it captures detailed skeletal data in the areas of interest with high accuracy.

To perform the expression identification task within Component One using the captured skeletal point data, various pattern recognition techniques were considered. Initially, three potential techniques were identified as being particularly suitable. According to the study [20] mentioned above, an LSTM model was utilized, and with linear addition, the author combined it with the CNN used to extract facial and body movements. Also, another potential technique we identified was the DTW algorithm according to the [15] study. Moreover, as the final technique for pattern recognition, we identified the Cross-Correlation technique.

<sup>1</sup><https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/holistic.md>

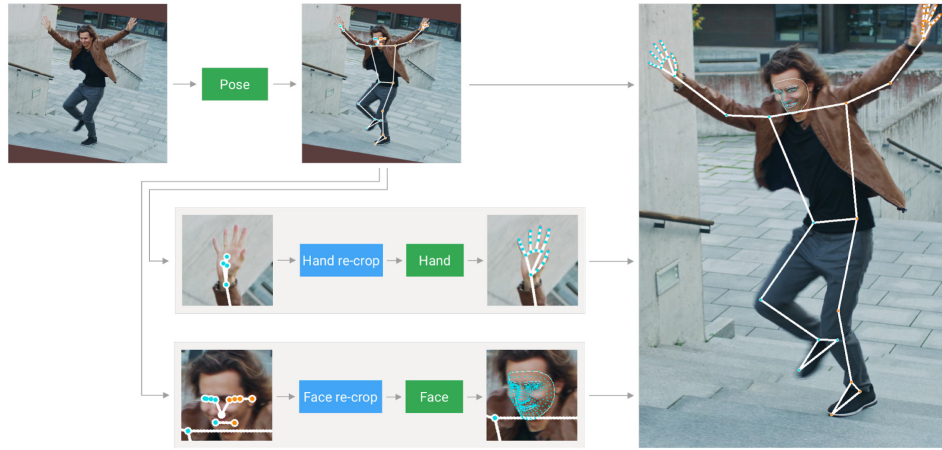


Figure 4.4: Skeletal points considered in the MediaPipe Holistic library

For component two, our primary focus was on exploring effective methods to create avatars, generate animations for different expressions, and deploy these avatars within a virtual environment. To develop the avatars, we utilized VRoid Studio, a powerful tool for designing custom 3D characters with detailed features suitable for a wide range of expressions. For the animation aspect, we incorporated Mixamo, an online platform that offers a vast library of pre-built animations. We selected and customized animations that best matched various expressions, such as high laugh, subtle laugh, surprise, and neutral, ensuring the avatars could convincingly express different feelings. Once the avatars and animations were ready, we integrated them into a virtual environment built using Unity. Unity served as the platform to bring together the animated avatars and the environment, enabling real-time expression representation and interaction. Our work mainly concentrated on seamless transitions between expression states and maintaining the avatar's consistency across different animations as shown in Figure 4.5. This approach provided a solid foundation for creating an immersive, expression-responsive virtual environment, essential for enhancing user engagement and realism in collaborative settings.

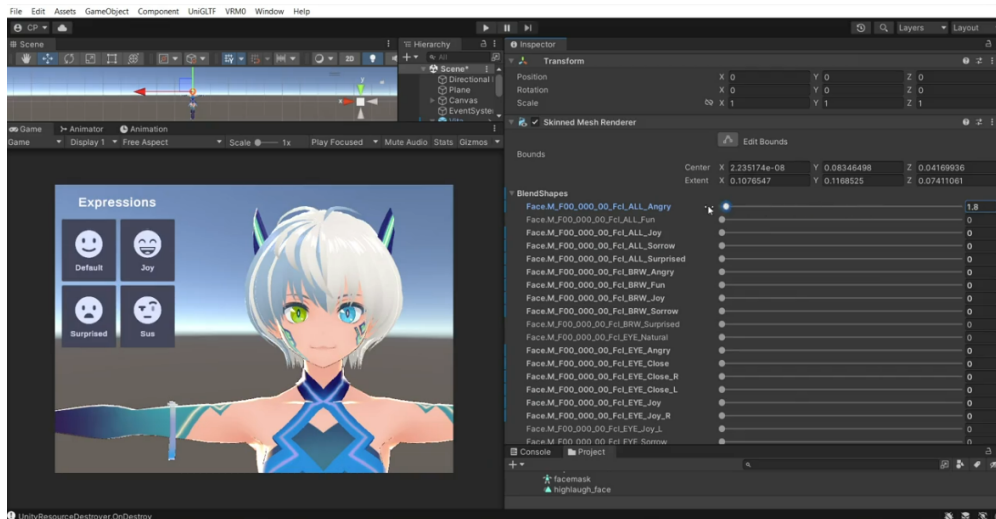


Figure 4.5: Expression Animations using VRoid Characters in Unity

Component Three, the system integration component, was designed to ensure seamless and near real-time communication between Component One and Component Two. The design emphasized the use of API calls for data exchange between the two components. Based on prevailing industry standards and the need for high performance, FastAPI was selected as the framework for building the inter-component communication layer.

Upon finalizing the system design and identifying the appropriate technologies for each component, the next focus was the development phase, focusing on implementing and integrating these components to create the functional system.

## **Identifying Facial Expressions and Body Gestures Related to Different Expressions**

In parallel to studying potential technologies to be used, research on different facial expressions and body gestures related to different types of expressions was conducted. In terms of facial expressions, the movements in the eyes, eyebrows, mouth, and cheeks were mainly focused on. Also, the head movements were considered as well. In terms of body gestures, the hands and shoulders were mainly noted. Once the expressions and gestures were recognized, a dummy dataset was created considering two main expressions, Happy and Confused, and a Neutral expression as well. This was created to make development easy, ensuring the components were working as intended. For happiness, two intensities were considered: High Laugh and Subtle Laugh.

To create the dummy dataset, we considered 5 participants. Participants were briefed on the research objectives and provided with examples of the four expressions high laugh, subtle laugh, confusion, and neutral as shown in 4.1 through explanations and short video clips based on our research as well as online resources, and asked each participant to perform the respective facial expressions and body gestures 30 times. A Python script was created to capture the skeletal points, and they were stored as numpy arrays. This process **addresses VirExp’s research question one**, which is "What specific facial expressions and body gestures convey distinct expressions, and what are the corresponding sequences of skeletal points associated with these gestures?"

## **Development**

During the development phase in iteration 1, we successfully implemented all three components and created an MVP. Here, under component 1, we created an LSTM-based pattern recognition model and a DTW-based pattern recognition algorithm. A model based on Cross-Correlation was not implemented because testing indicated issues in obtaining stable real-time data, especially on platforms like Google Colab, where latency and connection issues hindered performance. Local testing showed some improvement, but the technique ultimately lacked the robustness required for seamless real-time detection, especially in comparison to LSTM, which offers more reliable mapping in gesture recognition contexts.

To enhance the expressiveness of our 3D avatars, we leveraged mixamo animations, an

<b>Expression</b>	<b>Facial Expressions</b>	<b>Body Gestures</b>
Happy - High Laugh	Eyes mostly closed/open less, mostly raised cheeks, more wrinkles around the eye corners and widely opened mouth	Body extended, more upper body movements (forward and backward), more shoulder trembling and more forward head movements, hands held high
Happy - Subtle Laugh	More/normal eye openness, slightly raised cheeks, less wrinkles around the eye corners and mouth closed/simply showing teeth	Body extended, upper body remains relatively still, less shoulder movements and less head movements
Confused	Furrowed brow, raised eyebrows, a slightly open mouth or lips biting	Tilting the head, Scratching the head or touching face or mouth with hands
Neutral	Normal eye openness, relaxed face, mouth closed or slightly opened	Relaxed shoulders, no head or hand movements

Table 4.1: Facial Expressions and Body Gestures for Different Expressions

Adobe-owned platform for automated character rigging and motion capture animations, to generate and refine exaggerated expressions. Mixamo’s auto-rigging system streamlined skeletal animation for humanoid avatars, adhering to its requirements for neutral poses and clean mesh topology <sup>2 3</sup>. Its library of prerecorded animations (e.g. jumps, idle poses) served as a foundation, which we modified to amplify expression intensity (e.g. broader smiles, sharper frowns) while maintaining biomechanical plausibility <sup>3 4</sup>.

Research underscores that exaggerated animations improve expression recognizability in virtual environments, particularly where subtle expressions may be less perceptible<sup>4</sup>. For example, studies on behavioral realism demonstrate that avatars with exaggerated FACS-aligned expressions (e.g., increased puffing or raising of the cheeks) are perceived as more congruent with their expressions. We applied these principles by dynamically scaling Mixamo animations in Unity, adjusting parameters like ”energy” and ”overdrive” sliders to amplify key poses <sup>3</sup>. Challenges such as misaligned rotations in Unity (due to avatar configuration mismatches) were resolved by reprocessing animations in Blender to match the dimensions of the target skeleton.

The pipeline’s scalability was ensured through Mixamo’s modular animation system, allowing seamless integration of custom blendshapes for hyper-expressive states (e.g., surprise, anger). User testing confirmed that exaggerated animations enhanced perceived

<sup>2</sup><https://en.wikipedia.org/wiki/Mixamo>

<sup>3</sup><https://www.schoolofmotion.com/blog/4-ways-mixamo-makes-animation-easier>

<sup>4</sup><https://blogs.ulster.ac.uk/meganmccolm/2021/04/03/assignment-2/>



naturalness in virtual collaboration, aligning with findings from VR studies on motion clarity  
3 4 .

With these implementations, potential methods to **address research question two**, which is "In a real-time skeletal point sequence, how can we identify predefined facial expressions and body gesture patterns within near real-time and express them?" were identified.

## Demonstration

The demonstration phase of this research focused on evaluating the initial performance of the developed pattern recognition models (LSTM and DTW) within Component One of the system. For this initial evaluation, a group of ten IT-related undergraduate students, comprising five female and five male participants, were selected. Participants were within the age range of 24 and 25 years and had prior experience with virtual collaboration environments.

Prior to the commencement of the evaluation, each participant received a detailed explanation of the study's objectives and the specific facial expressions and body gestures associated with the four target expressions: high laugh, subtle laugh, confusion, and neutral. These associations were defined according to the criteria outlined in Table 4.1. To further facilitate understanding and accurate performance, suitable video examples for each of the four expressions were presented to the participants.

Following the explanation and visual examples, participants were asked to perform each of the four expressions. During their performance, the system captured the skeletal point data using Component One. The captured data was then processed by both the developed LSTM model and the DTW algorithm to predict the intended expression.

For each trial, participants were asked to complete a form (a sample of which is attached in the Appendix) to record two key pieces of information: the expression they intended to represent and the expression that was predicted by the model. This data collection method allowed for a direct comparison between the user's intended expression and the output of the pattern recognition models, providing valuable insights into their initial performance. This process was systematically carried out for both the LSTM model and the DTW algorithm to gather comparative performance data.

## Evaluation

Initially, the True Positive, False Positive, True Negative, and False Negative values were calculated for each gesture in order to calculate the model's performance using precision, recall, and F1 score. Table 4.2 shows the readings for each expression in the LSTM model, and Table 4.3 shows the precision, recall, and F1 scores for each expression in the LSTM model.

<b>Expression</b>	<b>Predicted: Neutral</b>	<b>Predicted: Subtle Laugh</b>	<b>Predicted: Higher Laugh</b>	<b>Predicted: Confused</b>
<b>Actual: Neutral</b>	4	6	0	0
<b>Actual: Subtle Laugh</b>	1	8	0	0
<b>Actual: Higher Laugh</b>	0	1	9	0
<b>Actual: Confused</b>	0	2	0	7

Table 4.2: Confusion Matrix of LSTM Model

<b>Expression</b>	<b>Neutral</b>	<b>Subtle Laugh</b>	<b>Higher Laugh</b>	<b>Confused</b>
<b>Precision</b>	0.8	0.47	1	1
<b>Recall</b>	0.4	0.8	0.9	0.7
<b>F1 Score</b>	0.53	0.59	0.95	0.82

Table 4.3: Precision, Recall and F1 Score of LSTM Model

Table 4.4 shows the readings for each expression in the DTW model, and Table 4.5 shows the precision, recall, and F1 scores for each expression in the DTW model.

<b>Expression</b>	<b>Predicted: Neutral</b>	<b>Predicted: Subtle Laugh</b>	<b>Predicted: Higher Laugh</b>	<b>Predicted: Confused</b>
<b>Actual: Neutral</b>	5	5	0	0
<b>Actual: Subtle Laugh</b>	1	9	0	0
<b>Actual: Higher Laugh</b>	0	0	10	0
<b>Actual: Confused</b>	0	2	0	8

Table 4.4: Confusion Matrix of DTW Model

<b>Expression</b>	<b>Neutral</b>	<b>Subtle Laugh</b>	<b>Higher Laugh</b>	<b>Confused</b>
<b>Precision</b>	0.83	0.56	1	1
<b>Recall</b>	0.5	0.9	1	0.8
<b>F1 Score</b>	0.62	0.69	1	0.89

Table 4.5: Precision, Recall and F1 Score of DTW Model

## **Communication**

The performance of the developed pattern recognition models, LSTM and DTW, was analyzed based on their F1 scores. While the DTW model generally exhibited higher F1 scores compared to the LSTM model, it was observed that the LSTM model provided faster predictions. These findings were considered key learnings for the subsequent iteration of the research. Also, the performance of the LSTM model tends to increase when the dataset size is increased. It was able to predict better when the dataset size is increased. In addition to the above learnings, a new technique, the Transformer model, was identified for the pattern recognition part in the pipeline. Moreover, the expression "Confusion" can be represented in so many ways, introducing a huge variation within one class.

### **4.2.3 Iteration 2 (November 2024 to April 2025)**

#### **Definition of Objectives for a Solution**

The objectives of the solution were not changed in iteration 2. However, objectives such as "Evaluate the effectiveness and performance of the system in facilitating user collaboration" is yet to be achieved, and solutions for other objectives are also to be refined in this iteration.

#### **Design and Development**

The second iteration of the design and development phase maintained the core system architecture established in the initial design, preserving the division into three primary components: expression capture and identification, expression representation, and system integration. However, this iteration focused on refining the existing technologies, exploring new potential technologies for performance enhancement, and continuing the detailed study of facial expressions and body gestures related to different expressions.

#### **New Technology Inclusion and Refinement of Existing Technologies**

Significant effort was dedicated to refining the existing LSTM model used for pattern recognition within the first component. A systematic hyperparameter tuning process was undertaken to optimize the model's performance. This involved exploring various combinations of hyperparameters, including learning rate, epoch values, and dropout ratios. The tuning process focused on monitoring key metrics such as validation accuracy and validation loss to identify the configuration that yielded the best performance. The accuracy and loss graphs illustrating the performance of the best-performing LSTM model during this hyperparameter tuning are presented in Figure 4.6 and Figure 4.7, respectively.

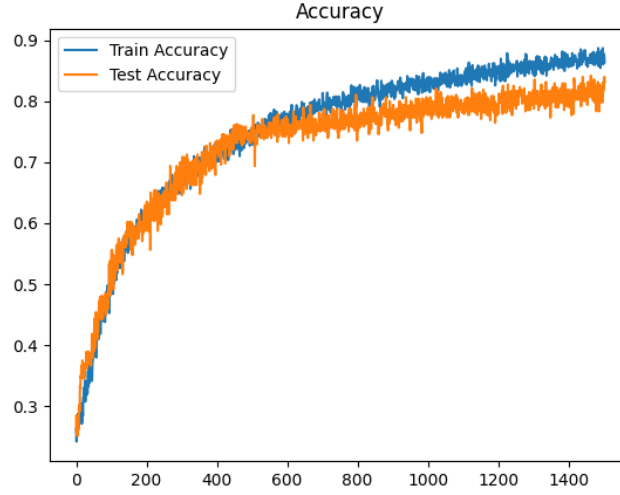


Figure 4.6: Accuracy Graph for the LSTM Model

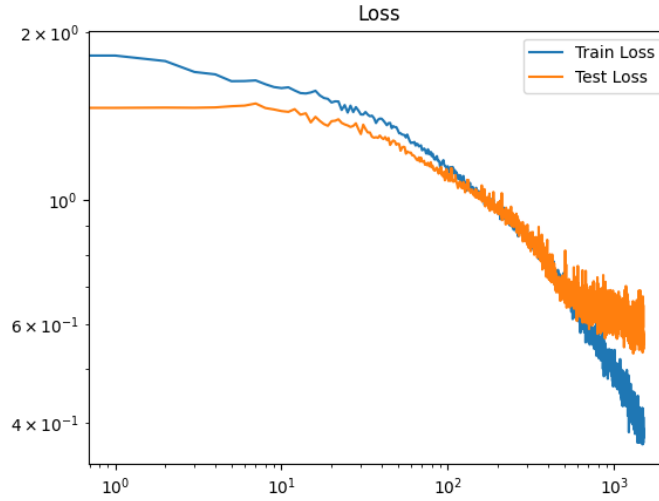


Figure 4.7: Loss Graph for the LSTM Model

Furthermore, a potential new technology for the pattern recognition task in Component One was identified and explored: the Transformer model. Recognizing the Transformer's capabilities in capturing long-range dependencies in sequential data, a Transformer model was implemented. This model was designed to predict the expression conveyed by the user's facial expressions and body gestures by comparing the captured skeletal point sequences to stored data arrays. Similar to the LSTM model, a comprehensive hyperparameter tuning process was conducted for the Transformer model. The best-performing hyperparameters were identified through the analysis of validation accuracy and validation loss graphs, which are presented in Figure 4.8 and Figure 4.9.

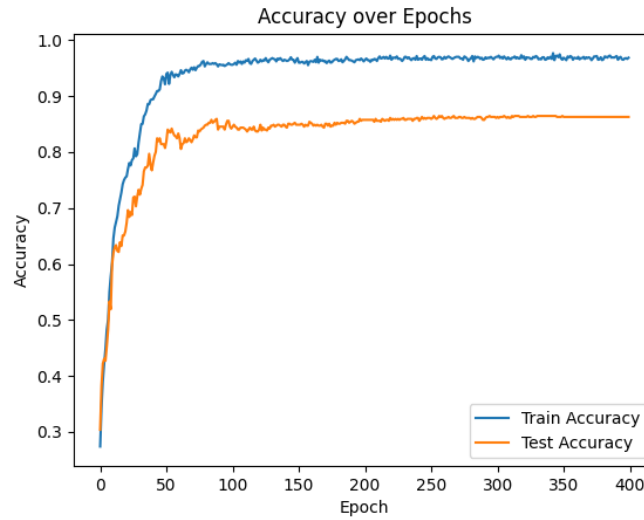


Figure 4.8: Accuracy Graph for the Transformer Model

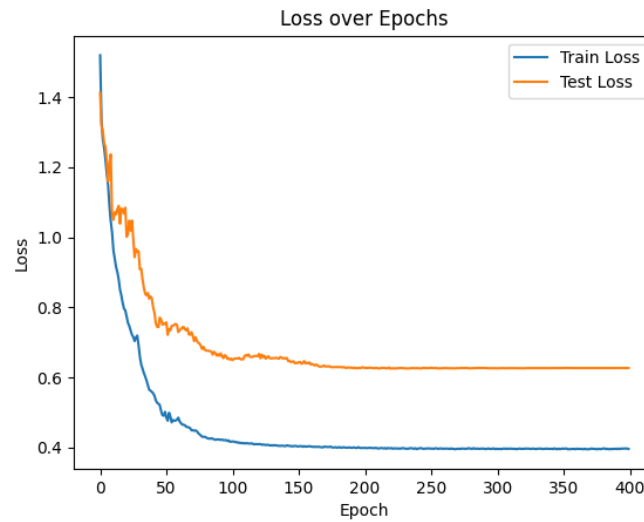


Figure 4.9: Loss Graph for the Transformer Model

## New Facial Expressions and Body Gestures Identified

Research was conducted on the nuances of facial expressions and body gestures associated with different expressions to refine the set of target expressions for the system. Based on this research and the goal of achieving accurate and distinct predictions, the expression of "confusion" was removed from the set. This decision was made because accurately recognizing confusion requires considering a wide array of subtle facial expressions and body gestures, leading to significant variation within the class, which could potentially result in inaccurate predictions.

In place of confusion, the expression of "surprise" was introduced into the set of target expressions. Surprise was chosen due to its relatively clear and distinct facial expressions

and body gestures compared to confusion, which is expected to contribute to more accurate recognition by the models.

## Dataset Creation

One of the key focuses of iteration 2 was to create a solid dataset to use for real-time expression detection. To create the dataset, we used data of 25 participants. Each participant performed every expression 10 times, and their skeletal points were captured and stored as numpy arrays. Then the dataset was expanded by flipping all 30 frames horizontally.

The same steps were followed as in iteration one to create the dataset. Participants were briefed on research objectives, and they were given a clear understanding of the relevant facial expressions and body gestures for each expression.

The facial expressions and body gestures related to High Laugh, Subtle Laugh, and neutral were the same as in iteration one, and specific facial expressions and upper-body movements that were identified and defined as characteristic of surprise for this research are represented in Table 4.6.

Expression	Facial Expressions	Body Gestures
Surprised	Raised eyebrows, wide-open eyes, widely open mouth/jaw dropped	Quick backward movement, sudden head or hand gestures (hand coming near mouth or forehead).

Table 4.6: Facial Expressions and Body Gestures for Different Expressions

The refined set of target expressions for the system in this iteration thus became High Laugh, Subtle Laugh, Surprise, and Neutral. This adjustment to the target expressions informed the ongoing hyperparameter tuning and evaluation of the LSTM and Transformer models, ensuring they were optimized for distinguishing between these specific expressions.

## Creation of Component Two - Expression Representation

To improve real-time expression sharing in multiplayer environments, we developed a networked Unity application using NGO and Unity Relay services, enabling synchronized avatar interactions across clients. The system captures user expressions via a webcam, processes them through a Python-based FAST API server, and transmits the data as JSON objects to Unity. These objects trigger corresponding facial animations on a VRoid character, ensuring remote players see the user’s expressions mirrored in real time as shown in Figure 4.10 and Figure 4.11 .



Figure 4.10: User in Multiplayer Environment



Figure 4.11: Representing expressions in Multiplayer Environment

To prevent unintended expression triggers, such as when the user is not actively facing the camera (e.g., talking to someone in the real world) we added a face orientation check. The system only detects and sends expressions when the user is directly facing the camera within a set angle threshold. This ensures the virtual avatar reflects only intentional expressions, avoiding confusion for other players.

Research highlights the importance of low-latency expression transmission for believable social presence in virtual spaces, particularly when subtle cues like smiles or frowns impact communication clarity. Our pipeline leverages MediaPipe’s holistic facial landmark

detection to map key expression parameters to VRoid blendshapes. Challenges such as network delays or desynchronized animations were mitigated by optimizing JSON payloads and implementing client-side prediction to smooth transitions between expression states. The integration of Unity Relay ensured scalable, secure peer-to-peer connectivity without dedicated servers, while the NGO handled state synchronization for avatar movements and expressions

## Demonstration

The demonstration process in the second iteration largely mirrored that of the first iteration, adapted to evaluate the refined models and the updated set of target expressions. The primary objective was to assess the performance of the enhanced pattern recognition models in identifying the expressions High Laugh, Subtle Laugh, Surprised, and Neutral.

A total of 30 participants were involved in the study, with their ages ranging from 18 to 35 years. The majority of participants (27, or 90%) fell within the 18–25 years age group, while the remaining 3 (10%) were between 26–35 years.

In terms of gender distribution, the sample was nearly balanced, with 15 male participants (50%) and 14 female participants (46.7%). One participant’s gender data was not specified.

None of the participants reported prior experience in similar research studies, ensuring unbiased responses.

Regarding familiarity with virtual collaboration tools and spaces—including platforms like Zoom, Google Meet, Microsoft Teams, and Metaverse environments (e.g., Decentraland, Sandbox, RecRoom, and PartySpace) the participants exhibited varying levels of experience:

- Medium familiarity: 18 participants (60%)
- High familiarity: 9 participants (30%)
- Low familiarity: 3 participants (10%)

This distribution suggests that most participants had moderate to high exposure to virtual collaboration, which may influence their perceptions and interactions in the study.

As before, participants were provided with a detailed explanation of the study’s purpose and the specific facial expressions and body gestures corresponding to the four target expressions. This explanation was informed by findings from the literature review and relevant online resources. Participants were shown visual examples to ensure a clear understanding of each expression.

Participants were then asked to perform each of the four target expressions. The system captured the skeletal point data in real-time using the MediaPipe Holistic library. The captured data was subsequently processed by the refined system to predict the intended expression.

To gather feedback on the system’s performance and the perceived accuracy of the expression recognition, participants were provided with a form (attached in the appendix)



to fill out after performing the expressions. More details about the features focused on in the evaluation will be discussed in the "Results and Evaluation" section.

## Evaluation

Details about the final evaluation is included in the "Results and Evaluation" section.

## Communication

Final communication points under interaction two will be discussed under the "Discussion" section.

## 4.3 Technical Explanation of System Implementation

### 4.3.1 Pattern Recognition Using the LSTM Model

LSTM networks are a specialized type of RNNs, particularly effective for processing sequential data, like the time-series keypoint data used for action recognition. Their core strength lies in their ability to learn long-range dependencies, understanding how events early in a sequence influence later events. This is achieved through an internal memory mechanism, called the cell state, and gating units (input, forget, output gates) that regulate the flow of information, allowing the network to remember relevant past information and forget irrelevant details, thus overcoming limitations like the vanishing gradient problem found in simpler RNNs. This makes LSTMs suitable for tasks like recognizing actions (laughter, surprise, neutral) from the sequence of 30 frames of keypoints extracted via MediaPipe.

The LSTM model implemented expected input data structured as sequences. Specifically, each input sample should have the shape (sequence\_length, number\_of\_features). The sequence\_length is set to 30, because it processes 30 consecutive frames at a time to understand an action. The number\_of\_features per frame is 1662, derived from concatenating the coordinates of keypoints for pose, face, left hand, and right hand detected by MediaPipe's Holistic solution. Therefore, the shape fed into the first LSTM layer during training or prediction is (batch\_size, 30, 1662).

The model architecture utilizes a stack of three LSTM layers. The first layer has 64 units and is configured with return\_sequences=True, meaning it outputs the hidden state for every one of the 30 time steps, passing the full sequence information onward. The second layer, with 128 units, also has return\_sequences=True, allowing the model to learn increasingly complex temporal patterns by processing the output sequence from the previous layer. The final LSTM layer has 64 units but uses return\_sequences=False. This instructs the layer to output only the hidden state corresponding to the very last time step (the 30th frame), effectively summarizing the information extracted from the entire sequence by the LSTM stack. All LSTM layers in this model use the tanh activation function internally.

To prevent overfitting, regularization techniques were incorporated. Within the LSTM layers themselves, dropout=0.2 was applied to the inputs, and recurrent\_dropout=0.2 is

applied to the recurrent connections (the connections feeding back within the LSTM cell). These techniques randomly deactivate a fraction of connections during training, forcing the network to learn more robust representations. Additionally, standalone Dropout(0.3) layers are inserted after the LSTM stack and between the Dense layers to further mitigate overfitting.

Following the LSTM layers that process the temporal dynamics, standard fully connected Dense layers were added (Dense(64, ...) and Dense(32, ...)) using the ReLU activation function. These layers take the summarized sequence representation from the final LSTM layer and perform further nonlinear transformations, helping to map the learned temporal features to the classification task.

The final layer is a Dense layer with a number of units equal to the number of action classes (`actions.shape[0]`, which is 4 in this case). It uses the softmax activation function. Softmax converts the raw output scores from the network into a probability distribution across the different action classes ('highLaugh', 'subtleLaugh', 'surprised', 'neutral'), ensuring the probabilities sum to 1 and indicating the model's confidence for each potential action.

During training, the model is compiled using the Adam optimizer and the categorical\_crossentropy loss function, appropriate for multi-class classification where labels are one-hot encoded. Accuracy (categorical\_accuracy) is tracked as a performance metric. The training process also employs EarlyStopping by monitoring the validation loss, which helps prevent overfitting by halting training when performance on unseen data stops improving and restoring the model weights from the best-performing epoch. A graphical representation of the LSTM layers is shown in Figure 4.12.

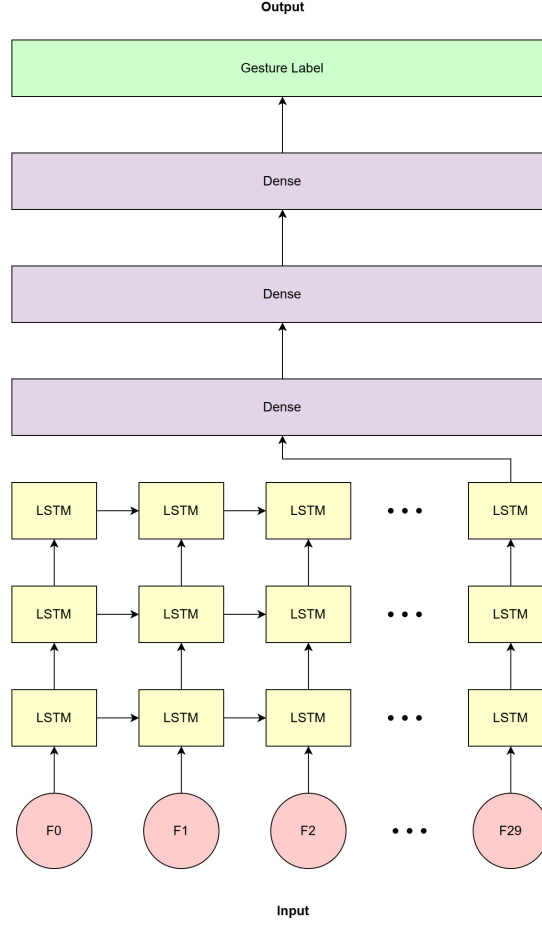


Figure 4.12: Layers of the LSTM Model

### 4.3.2 Pattern Recognition Using the DTW Algorithm

DTW is an algorithm designed to measure the similarity between two temporal sequences that may vary in speed or timing. Unlike methods that require sequences to be of the same length and perfectly aligned, DTW finds an optimal non-linear alignment between points in the two sequences, calculating a cumulative distance or cost associated with this alignment. This makes it suitable for comparing performed gestures, which can naturally vary in duration and execution speed even when representing the same action.

The input data for the DTW process consists of sequences of keypoints captured frame-by-frame using MediaPipe’s Holistic solution. Each frame is represented by a vector of 1662 features, derived from concatenating the coordinates x, y, z, and sometimes visibility of pose, face, left-hand, and right-hand landmarks. However, in this study, sequences of a defined length were considered, 30 frames in this case, to represent a complete gesture instance.

The core mechanism DTW algorithm used is template matching. Instead of training a model like an LSTM, it relies on a pre-existing dataset of recorded gesture examples (templates). It loads multiple (480 per action) template sequences for each gesture class ('highLaugh', 'subtleLaugh', 'surprised', 'neutral') from stored .npy files. These stored sequences serve as the reference templates against which new, incoming gestures will be

compared.

During the real-time recognition phase, the system continuously captures frames from the camera, extracts keypoints, and collects them into a sequence. Once this collected sequence reaches the target length (30 frames), it is compared against every single template sequence loaded previously for all gesture classes. The comparison is performed using a DTW distance calculation to find the similarity score between the collected sequence and each template.

Classification is achieved by finding the template sequence that yields the minimum DTW distance to the input sequence. The gesture label ('highLaugh', 'subtleLaugh', etc.) associated with this best-matching template is considered the most likely candidate for the performed gesture. A threshold (`dtw_threshold = 200`) is then applied to this minimum distance. If the distance is below the threshold, the corresponding gesture is recognized and displayed; otherwise, if the minimum distance is too high (indicating even the best match isn't very similar), the input is classified as 'Gesture Not Recognized'. This threshold helps reject sequences that don't closely resemble any known templates.

However, the DTW-based solution was not selected to build the final pipeline because its latency to detect the expression (45-50 seconds) was not suitable for this solution that operates within near-real time.

### 4.3.3 Pattern Recognition Using the Transformer Model

This model uses a different approach compared to LSTMs or DTW, leveraging the Transformer architecture, originally known for its success in natural language processing, but adapted here for sequence data derived from body keypoints.

The primary input to the model is sequences of keypoints extracted using MediaPipe, representing actions like 'highLaugh', 'subtleLaugh', 'surprised', or 'neutral'. Each sequence consists of 30 time steps (frames), and each time step contains a flattened vector of 1662 keypoint features (`sequence_length = 30`, `feature_dim = 1662`). The initial step within the model involves a Dense layer (projection) which takes this input (`batch_size, 30, 1662`) and projects each time step's 1662 features into a higher-dimensional embedding space of size 384 (`embed_dim = 384`). This results in a sequence of embeddings with the shape (`batch_size, 30, 384`), which is then processed by the core Transformer layers.

The heart of the model is a stack of four Transformer Encoder layers. Each layer performs two main operations. First, it uses a Multi-Head Self-Attention mechanism. Unlike LSTMs that pass information sequentially, self-attention allows each frame's embedding in the sequence to look at and weigh the importance of all other frame embeddings (including itself) within the entire 30-frame sequence simultaneously. This helps capture complex relationships and context across the whole action sequence, regardless of distance. The "Multi-Head" aspect means this attention process is done in parallel across 8 different "heads", each potentially focusing on different types of relationships, and their results are combined. After attention, a residual connection (adding the input of the sub-layer to its output) and layer normalization are applied to aid training.

The second main operation within each encoder layer is a position-wise FFN. This consists of two dense layers: the first expands the dimension to `hidden_dim = 768` with a ReLU activation, and the second projects it back down to the `embed_dim = 384`. This network processes each sequence position (time step) independently but identically. Again, a residual connection and layer normalization follow this sub-layer. A dropout of 0.2 was applied within both the attention and feed-forward sub-layers during training as a regularization technique to prevent overfitting.

After the input sequence has passed through all four encoder layers, the resulting sequence of processed embeddings needs to be converted into a format suitable for classification. This is done using global average pooling, which averages the embeddings across the time dimension (the 30 frames). This produces a single, fixed-size vector representation for the entire input sequence.

Finally, this pooled vector representation is fed into a standard Dense classification layer with a softmax activation function. This layer outputs a probability distribution over the four action classes, indicating the model’s prediction for the input sequence. The model was trained using the Adam optimizer with a Cosine Decay learning rate schedule and gradient clipping, minimizing the categorical cross-entropy loss (with label smoothing applied). In summary, the Transformer model processes the entire sequence of keypoints in parallel using self-attention to capture relationships across frames, followed by pooling and a final classification layer.

However, the Transformer model was not selected to build the pipeline due to its high resource consumption and because it is challenging to run the model in a normal setup.

#### 4.3.4 Expression Representation Component

The Unity multiplayer framework was implemented using a combination of scripts to manage networking, player synchronization, and game mechanics, alongside a sophisticated avatar system for immersive interactions. The Relay Manager facilitated the integration of Unity Relay services, enabling secure peer-to-peer connections through relay servers. For instance, the `SetupRelay()` method created a relay allocation and generated a join code, while `JoinRelay()` allowed clients to connect using this code. This ensured seamless multiplayer functionality even behind NATs, with data such as IP addresses, ports, and allocation IDs being synchronized via `UnityTransport`.

Player management was handled by `Players Manager`, which tracked the number of connected players using a `Network Variable`. Server-side callbacks incremented or decremented the player count when clients connected or disconnected. Meanwhile, `Player ID Logger` displayed each player’s unique `Local Client Id` on their UI, ensuring players could identify their own instance. For example, the `Start()` method logged the ID and updated a `Text Mesh ProUGUI` element, providing immediate visual feedback.

The avatar system was built using customizable 3D models from `VRoid Studio`, which were rigged with Unity’s `Humanoid Avatar` system for consistent skeletal animation.

Animation states were managed by a BlendTree-based animator controller, allowing smooth transitions between expressions such as "surprise," triggered by facial detection inputs like raised eyebrows. Real-time synchronization was achieved using Unity Netcode, which replicated avatar states (e.g., movement, expressions) across all connected clients, while Discord handled external voice chat to offload bandwidth from Unity's networking.

Player movement and animation were controlled by Player Control Authoritative, which used Client Network Transform for synchronized positioning. The script processed input to transition between states like Walk, Run, and Idle, updating animations via Network Variable. For instance, holding the Shift key triggered the Run state, while releasing it reverted to Walk. The ServerRpc method ensured state changes were replicated across all clients, maintaining consistency in the virtual environment.

UI interactions were managed by UIManager.cs, which provided buttons for hosting, joining, and starting servers. The script integrated with RelayManager to handle relay-based connections when enabled, as seen in the host and client setup logic. Additionally, it displayed real-time player counts by referencing PlayersManager.Instance.PlayersInGame. The PlayerHud.cs script further enhanced UI by assigning and displaying player names (e.g., "Player 1") using a Network Variable, which synchronized text data across the network.

Object spawning was delegated to Spawner Control, which utilized a network object pool to instantiate prefabs server-side. For example, the SpawnObjects() method created multiple instances at random positions, optimizing performance through pooling. Camera switching, implemented in Cam Switch, allowed toggling between camera groups using keyboard inputs (e.g., pressing 'C' or 'V'), demonstrating modular control over view perspectives.

Finally, Network String and Relay Join Data provided essential data structures. The former serialized strings for network transmission using FixedString32Bytes, while the latter stored relay connection details like join codes and allocation IDs. Together, these scripts and systems formed a robust multiplayer framework, enabling synchronized gameplay, dynamic UI updates, and efficient resource management, while the avatar system enhanced user immersion through expressive, real-time animations.

Additionally, a Discord channel was used in the process to enable real-time voice communication during virtual interactions, providing a seamless and low-latency solution for participants. The choice of Discord was justified by its widespread adoption, ease of use, and robust voice chat capabilities, which are essential for maintaining clear and uninterrupted communication during collaborative tasks <sup>5</sup>. Additionally, Discord's cross-platform compatibility ensures accessibility across various devices, while its low resource consumption minimizes performance disruptions during system operation.

### 4.3.5 System Integration Component

The integration between the Python-based detection module and the Unity virtual environment was implemented using FastAPI, a lightweight yet high-performance web

---

<sup>5</sup><https://discord.com/features>

framework for building APIs in Python. This setup facilitated seamless communication between the two systems, with expression labels (e.g., "high laugh", "surprise") transmitted as JSON packets over a local HTTP connection. The choice of FastAPI was deliberate due to its asynchronous request handling, which minimized overhead and ensured low-latency data transfer between the real-time detection system and the Unity client [27].

Given that most facial expressions and gestures evolve over multiple frames rather than instantaneously, the Python detection module was configured to process every third frame (10 FPS) instead of analyzing all 30 FPS from the webcam. Empirical testing confirmed that this reduced computational load by 66% without significantly degrading detection accuracy [28]. This approach was particularly effective because facial expressions typically persist for 200–500ms [29], making sub-sampling feasible without losing critical temporal information.

The Figure 4.13 below shows the high-level architecture of VirExp solution that is presented by this study.

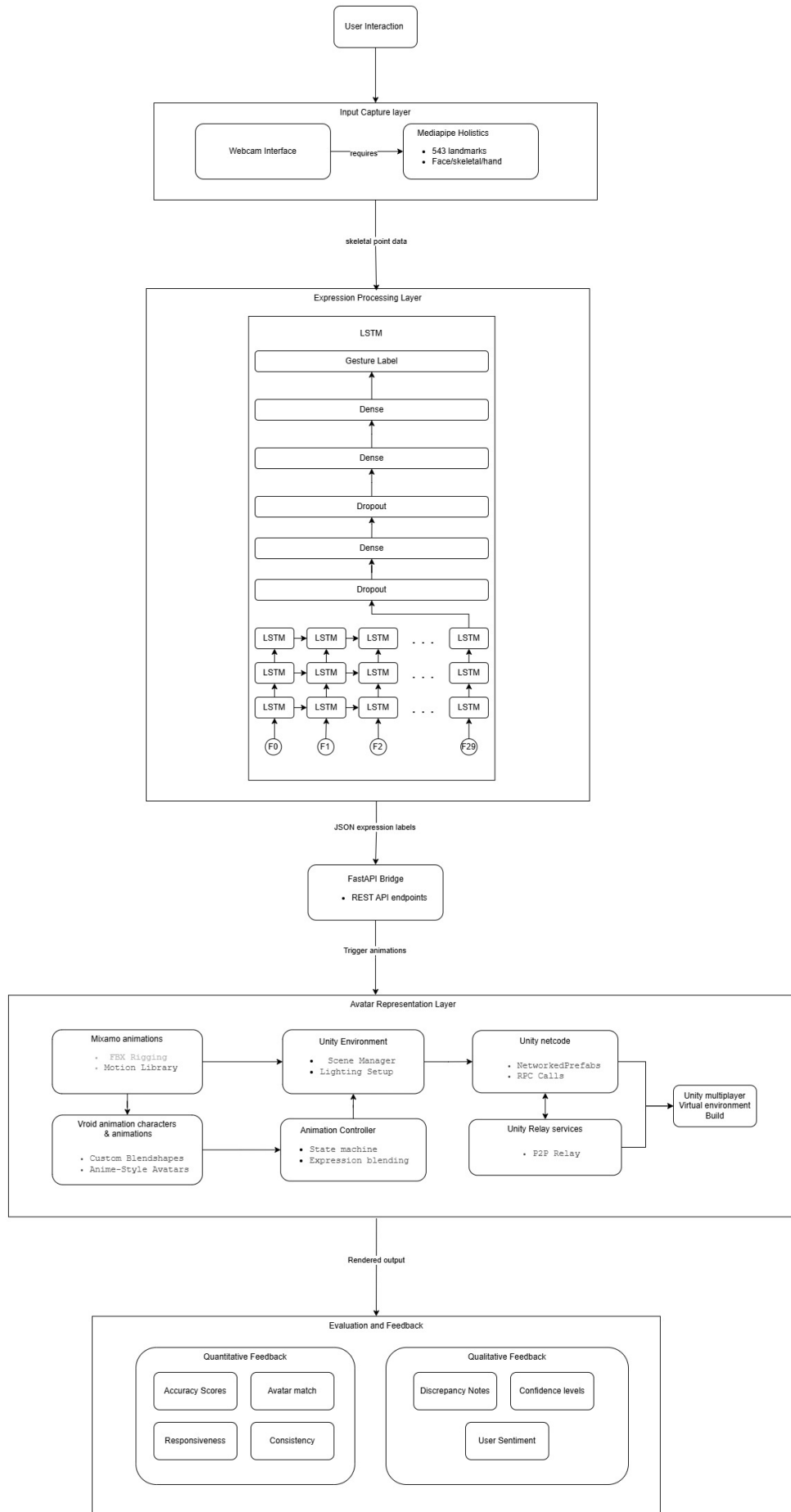


Figure 4.13: High-Level Architecture Diagram of VirExp



# Chapter 5

## Results and Evaluation

### 5.1 Presentation of Findings

#### Process Evaluation Scope

This study focuses specifically on evaluating four archetypal expressions (**High Laugh**, **Subtle Laugh**, **Surprise**, and **Neutral**) rather than assessing a comprehensive range of expressions.

The **Process Validation** examines whether our detection-to-representation pipeline functions correctly for these fundamental expressions before expanding to more complex expressions. For **Controlled Assessment**, these four expressions were selected because they represent distinct facial action unit patterns and cover a spectrum of intensity from subtle to exaggerated.

The **Methodological Precision** of limiting variables enables clearer analysis of potential breakdowns in either the expression detection phase, the avatar representation phase, or the overall user experience regarding real-time performance, consistent performance, and enhancement of collaborative experience. This constrained evaluation serves as the essential first step in validating our technical and theoretical framework before broader implementation, with future studies planned to incorporate additional expressions once this foundational process demonstrates reliability.

#### Study Structure

Participants will engage in four structured trials - one for each target expression (**High Laugh**, **Subtle Laugh**, **Surprise**, and **Neutral**). This focused approach enables systematic evaluation of our detection-to-representation pipeline for these fundamental expressions that represent distinct facial action patterns across a complete intensity spectrum.

We maintain rigorous experimental control through isolation of each expression type in dedicated trials, minimizing cognitive load for participants while enabling clear identification of system strengths and weaknesses per expression category. This structure also establishes baseline performance metrics essential before expanding to more complex expression ranges in future research.

## **Trial Format**

Each trial follows a consistent three-phase protocol: First, participants perform one specific expression. Second, our system detects and translates it to avatar animation. Third, participants evaluate the accuracy of both detection and representation relative to their intention. This structured, iterative testing approach enables precise identification of needed improvements in either technical processing or visual representation components.

## **Expression Detection and Animation Performance**

For **High Laugh**, most participants reported satisfactory detection and animation, though some noted minor discrepancies including occasional confusion with Surprise or recognition delays. Several participants observed the system relied more on upper body movements than facial details for detection.

The **Subtle Laugh** expression presented detection challenges, with multiple reports of misclassification as Neutral or Surprise. Participants noted particular inconsistencies when the laugh occurred without accompanying sound, suggesting audio cues may influence detection accuracy.

**Surprise** showed variable detection accuracy, with some confusion occurring between Surprise and High Laugh expressions. In several cases, the system failed to recognize the expression entirely, defaulting to Neutral.

**Neutral** expressions demonstrated the most reliable performance, with most participants reporting accurate detection and animation. The primary issue involved rare instances where Neutral was mistaken for Subtle Laugh.

## **System Responsiveness and Consistency**

Evaluation of system responsiveness revealed generally positive feedback, though participants noted slight recognition delays particularly for Subtle Laugh expressions. Consistency during prolonged use received favorable reports, with minimal instances of interruptions or performance degradation observed across testing sessions.

## **User Experience and Adoption Potential**

Participant feedback indicated the system enhanced collaborative experience for most users, who highlighted its intuitive design and potential for improving virtual communication. While a majority expressed willingness to adopt such a system in future sessions, a smaller contingent remained uncertain, identifying specific areas requiring refinement for optimal utility.

## **5.2 Data Analysis**

To evaluate the performance and reliability of the developed expression recognition system, a comprehensive analysis was conducted using a multi-class classification approach. The

model was tested against four distinct expressions High Laugh, Subtle Laugh, Surprised, and Neutral capturing both expressive and subtle expression variations in virtual collaborative environments. Each expression was assessed using key evaluation metrics such as precision, recall, F1-score, and accuracy, which provided a detailed understanding of the system’s strengths and limitations in real-time expression detection. Furthermore, a confusion matrix was generated to visualize the distribution of correct and incorrect predictions across all expression classes. This matrix served as a valuable tool for identifying common misclassification patterns and evaluating the model’s discriminative capability between similar expressions. The results obtained offer deep insight into the effectiveness of the system and form the foundation for discussing areas of improvement and potential for future enhancement.

## **Expression Recognition**

The overall performance of the expression recognition system across all four expression classes, High Laugh, Subtle Laugh, Surprised, and Neutral—was further analyzed using a confusion matrix. Out of 30 instances labeled as High Laugh, the system correctly classified 28 of them. It misclassified 1 as Surprised and 1 as Neutral, with no confusion with Subtle Laugh. For Subtle Laugh, 26 instances were correctly predicted, while 1 was misclassified as Surprised and 3 as Neutral, again with no confusion with High Laugh. In the case of Surprised, 27 instances were accurately identified, with 1 misclassified as High Laugh and 2 as Neutral. The Neutral expression showed the strongest individual performance with 29 correctly classified instances and only 1 being misclassified as Subtle Laugh.

This confusion matrix highlights that most errors occurred between Subtle Laugh, Surprised, and Neutral, suggesting some overlapping features or expression ambiguity in these categories. Notably, the classifier avoided severe cross-category confusion (e.g., no High Laugh being misclassified as Subtle Laugh and vice versa), indicating strong discriminative capability among highly distinct expressions. Overall, out of 120 total instances, 110 were classified correctly, reflecting a total classification accuracy of 91.67%, and affirming the model’s effectiveness in multi-class expression recognition.

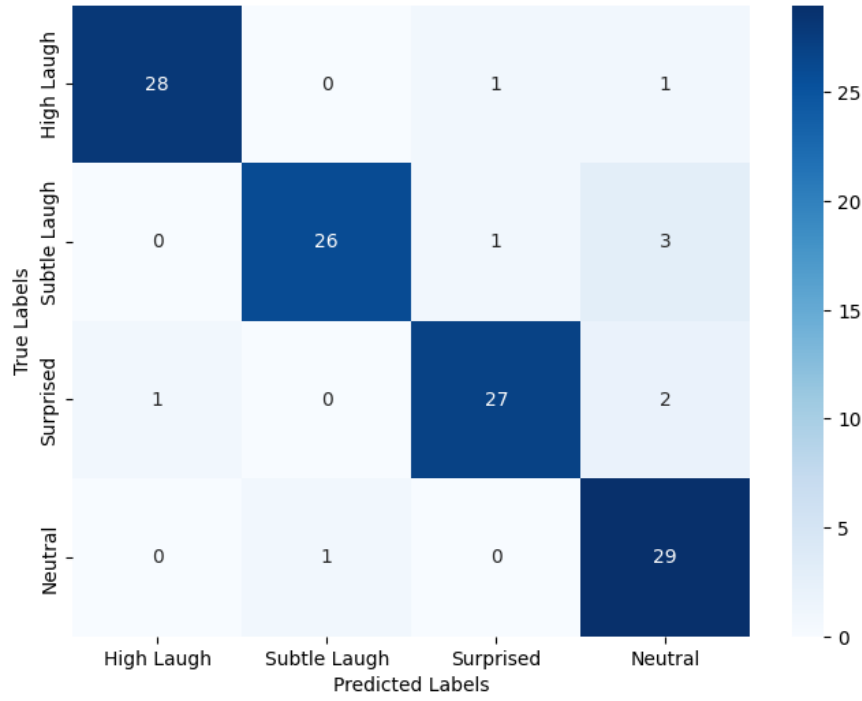


Figure 5.1: Confusion Matrix on the Overall Expression Recognition

### High Laugh Recognition

During the evaluation phase of the system, the recognition of the High Laugh expression demonstrated a high level of performance across key classification metrics. The model correctly identified 28 true positive cases, with only 1 instance incorrectly classified as a high laugh (false positive) and 2 actual high laugh expressions missed (false negatives). The model also successfully rejected 89 non-high laugh cases (true negatives). These results yielded a precision of approximately 96.55%, indicating that the majority of the model's high laugh predictions were accurate. The recall stood at 93.33%, reflecting the model's ability to correctly detect most high laugh instances from the dataset. The F1-score, which balances precision and recall, was 94.91%, suggesting strong overall performance in both identifying true positives and avoiding false results. Additionally, the model achieved a high accuracy of 97.5%, reinforcing its effectiveness in correctly classifying both positive and negative cases. These outcomes indicate that the system is highly reliable in detecting exaggerated high laugh expressions, making it suitable for real-time expression recognition applications in virtual collaborative environments.

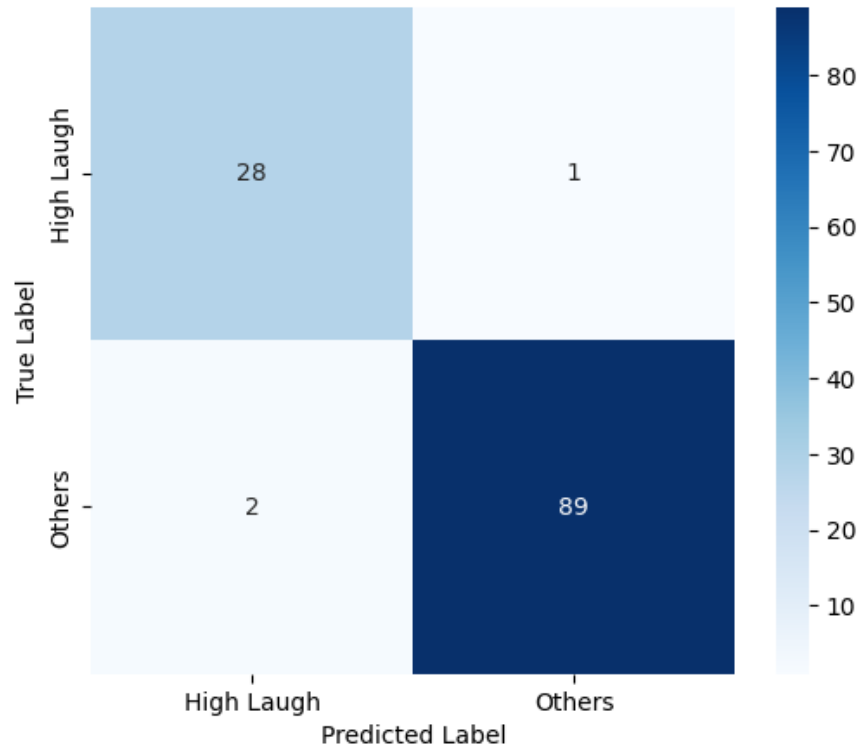


Figure 5.2: Confusion Matrix on the "High Laugh" Expression Recognition

### Subtle Laugh Recognition

In evaluating the system's ability to recognize the Subtle Laugh expression, the results reflected solid classification performance with minor room for improvement. The model achieved 26 true positives, correctly identifying subtle laugh expressions in the majority of relevant instances. There was only 1 false positive, indicating a very low rate of incorrect subtle laugh detections. However, the system missed 4 true subtle laughs (false negatives), which slightly affected its recall. Additionally, 89 true negatives were recorded, showing consistent reliability in correctly identifying non-subtle laugh cases. The precision for this expression was approximately 96.30%, suggesting that nearly all predictions made as subtle laughs were correct. The recall was 86.67%, indicating that the system was able to detect most but not all subtle laugh expressions. The F1-score, which combines both precision and recall, was calculated to be 91.23%, reflecting an overall strong performance. The system also maintained a high accuracy of 95.83%, demonstrating its general effectiveness. These metrics show that the system performs well in recognizing subtle laugh expressions, although improving recall could further enhance its sensitivity to less intense expressions.

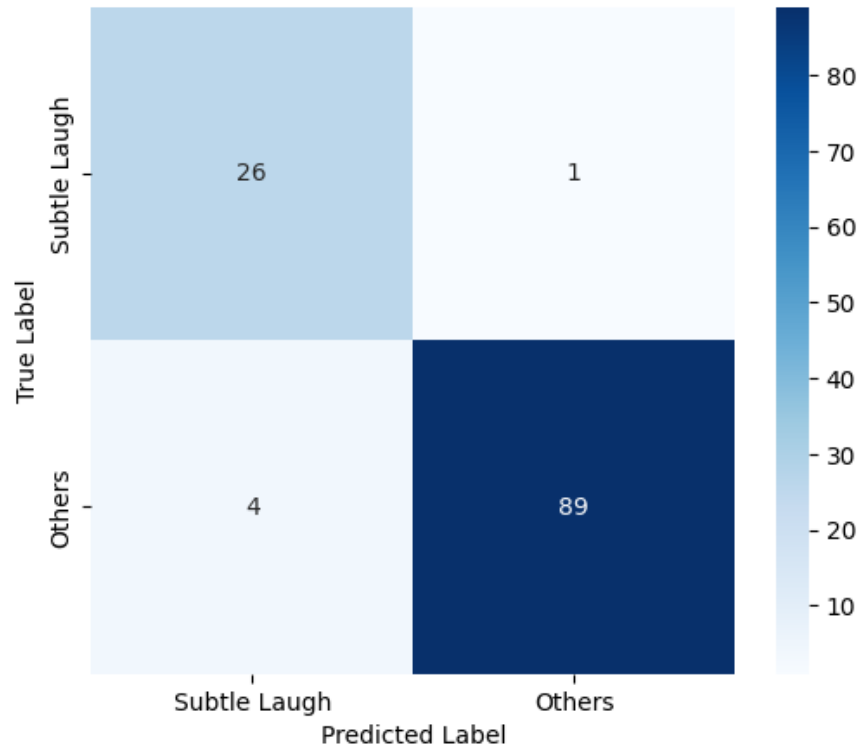


Figure 5.3: Confusion Matrix on the "Subtle Laugh" Expression Recognition

### Surprised Recognition

The system's performance in detecting the Surprised expression also demonstrated high reliability and consistency. The model successfully identified 27 true positives, correctly classifying the majority of surprised expressions. There were 2 false positives, where non-surprised instances were incorrectly labeled as surprised, and 3 false negatives, where actual surprised expressions were missed. The model correctly rejected 88 non-surprised cases (true negatives), maintaining its overall robustness. The precision for this expression was 93.10%, indicating that most surprised predictions were indeed correct. The recall reached 90.00%, signifying that the model detected a substantial portion of actual surprised expressions. The F1-score a harmonic mean of precision and recall stood at 91.53%, confirming the model's balanced performance in both identifying and distinguishing the expression accurately. With an overall accuracy of 95.83%, the system continues to show strong general classification ability. These results affirm that the model is effective in recognizing surprised expressions, with a relatively low error rate and strong predictive power across various cases.

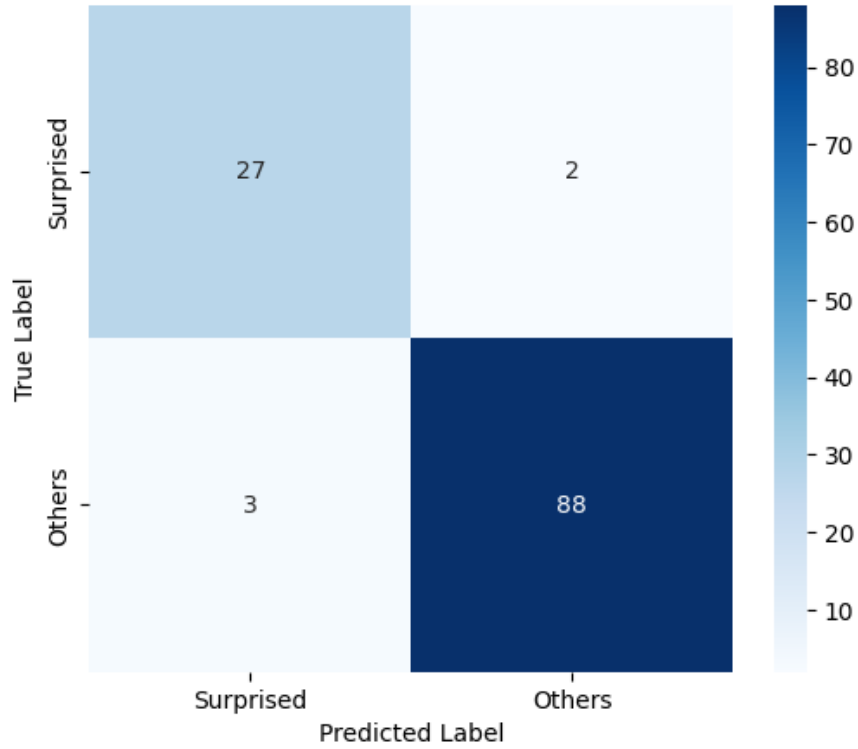


Figure 5.4: Confusion Matrix on the "Surprised" Expression Recognition

### Neutral Recognition

When assessing the system's performance in recognizing the Neutral expression, the results revealed both strengths and some areas that could benefit from refinement. The model achieved a high number of true positives (29), successfully identifying nearly all neutral expressions. It only missed 1 actual neutral case (false negative), resulting in a strong recall of 96.67%, which indicates high sensitivity to this expression. However, the model also produced 6 false positives, where other expressions were incorrectly classified as neutral, slightly lowering the precision to 82.86%. Despite this, the model correctly identified 84 non-neutral expressions (true negatives), contributing to an overall accuracy of 94.17%. The F1-score for neutral detection was 89.23%, balancing the high recall with the comparatively lower precision. These results suggest that while the system is very effective in detecting neutral expressions, it has a tendency to over-predict this category in some cases. Fine-tuning the model to reduce false positives would enhance its overall reliability in distinguishing neutral states from other expressions.

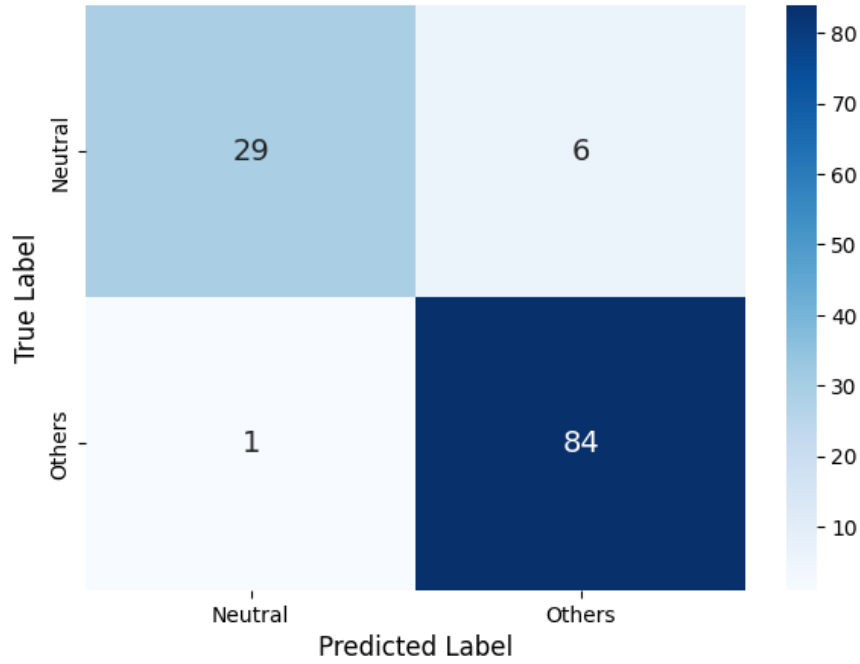


Figure 5.5: Confusion Matrix on the "Neutral" Expression Recognition

The evaluation of the expression recognition system across four target expressions—High Laugh, Subtle Laugh, Surprised, and Neutral reveals consistently high overall performance, with each expression yielding unique strengths and challenges. High Laugh demonstrated the strongest results, with an F1-score of 94.91%, precision of 96.55%, and a high accuracy of 97.5%, indicating excellent model performance in detecting exaggerated expressions. Subtle Laugh maintained similarly high precision (96.30%) but showed a slight drop in recall (86.67%), reflecting some difficulty in detecting softer expressions, yet still achieving an impressive F1-score of 91.23%. The Surprised expression also scored well across all metrics, with a balanced F1-score of 91.53%, precision of 93.10%, and recall of 90.00%, suggesting robust detection of distinct and high-energy expression patterns. Neutral, while achieving the highest recall (96.67%), had the lowest precision (82.86%) due to a relatively higher number of false positives, resulting in a slightly lower F1-score of 89.23%. Overall, the model exhibits strong accuracy and reliability, particularly in identifying expressions, with room for refinement in distinguishing more subtle or baseline expressions like Neutral.

### Expression Representation

The confusion matrix for the system's ability to recognize and represent expressions High Laugh, Subtle Laugh, Surprised, and Neutral in avatars within the virtual environment demonstrates the system's strong performance in most cases, with a few areas for improvement. Out of the 30 instances of High Laugh, the system correctly identified all 30 (True Positives), with no misclassifications. For Subtle Laugh, the system accurately recognized 25 out of 30 instances, but misclassified 4 instances as Neutral and 1 instance as Surprised. In the case of Surprised, the system correctly identified 28 of the 30 occurrences, with 2 instances missed as false negatives, while no instances were misclassified as other



expressions. Lastly, for Neutral, the system successfully identified 29 neutral expressions, but mistakenly labeled 1 instance as Subtle Laugh. The total number of evaluations across all expressions amounted to 120 instances, with a majority of correct classifications, indicating that the system performs well in recognizing these expressions, though improvements could be made in handling subtle expression nuances and reducing misclassifications between similar expressions like Subtle Laugh and Neutral.

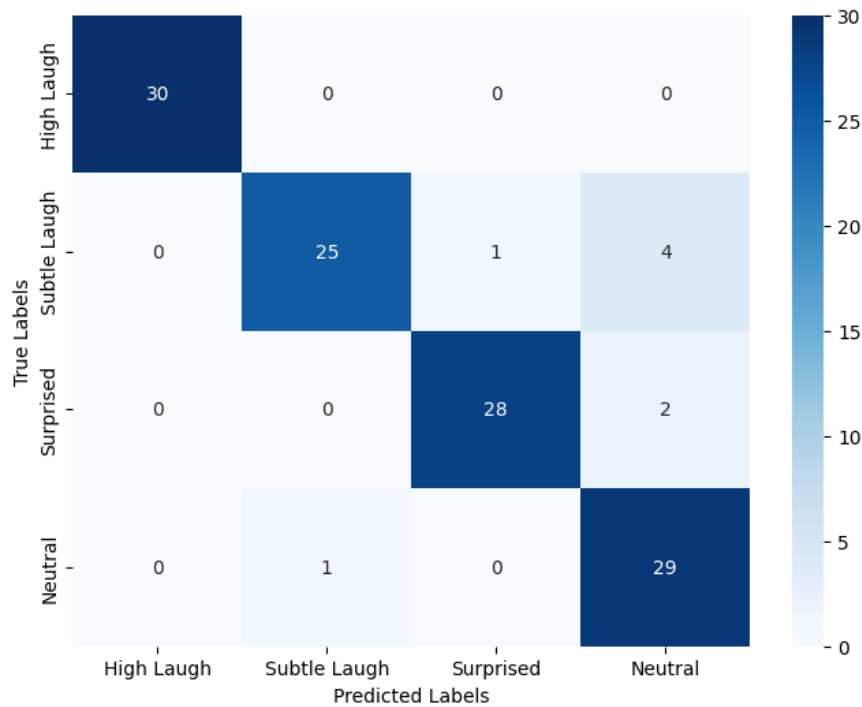


Figure 5.6: Confusion Matrix on the Overall Expression Representation

### High Laugh Representation

During the evaluation of the system’s capability to recognize and accurately represent the ”High Laugh” expression through avatars in a virtual environment, the results demonstrated exceptional performance across all standard classification metrics. The system achieved a True Positive count of 30, correctly identifying all instances where a high laugh expression was present. Moreover, there were zero False Positives and zero False Negatives, indicating that the system did not mistakenly classify other expressions as high laugh, nor did it miss any actual occurrences of it. With a True Negative count of 90, the system accurately recognized non-laugh expressions as such. Consequently, the precision, recall, F1-score, and accuracy all achieved the perfect score of 1.0. These results affirm the system’s outstanding ability to not only detect but also faithfully represent the ”High Laugh” expression through avatar animations, contributing to an immersive and responsive virtual environment.

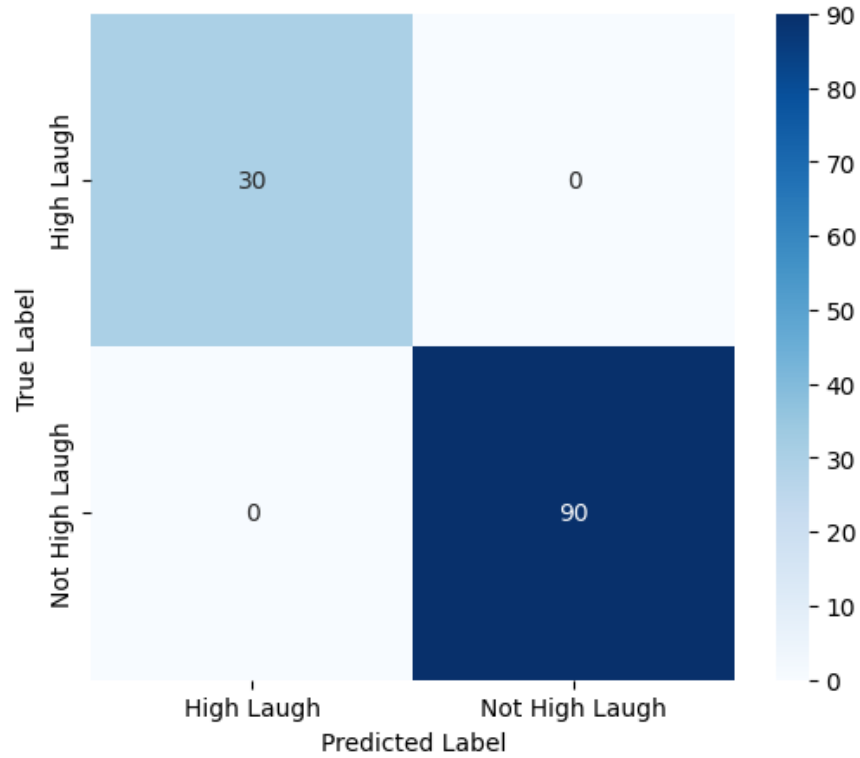


Figure 5.7: Confusion Matrix on the "High Laugh" Expression Representation

### Subtle Laugh Representation

In evaluating the system's performance in recognizing and representing the Subtle Laugh expression through avatars in the virtual environment, the results indicate a high level of accuracy with minor limitations. Out of all evaluated instances, the system correctly identified 25 occurrences of subtle laughter (True Positives), while only 1 instance was incorrectly classified as a subtle laugh when it was not (False Positive). However, the system failed to detect 5 actual occurrences of subtle laughter (False Negatives), suggesting room for improvement in capturing more nuanced expressions. With 89 correct rejections of non-subtle laugh expressions (True Negatives), the overall accuracy stood at 0.95, indicating solid reliability. The precision was measured at 0.96, reflecting the system's strong ability to avoid false alarms. The recall was slightly lower at 0.83, highlighting a need to enhance sensitivity to subtle expressions. The F1-score, a harmonic mean of precision and recall, was 0.89, suggesting a well-balanced but improvable performance. Overall, the system demonstrates effective, though not flawless, recognition and expressive rendering of subtle laughter in virtual avatars, supporting believable and expressively rich interactions in immersive environments.

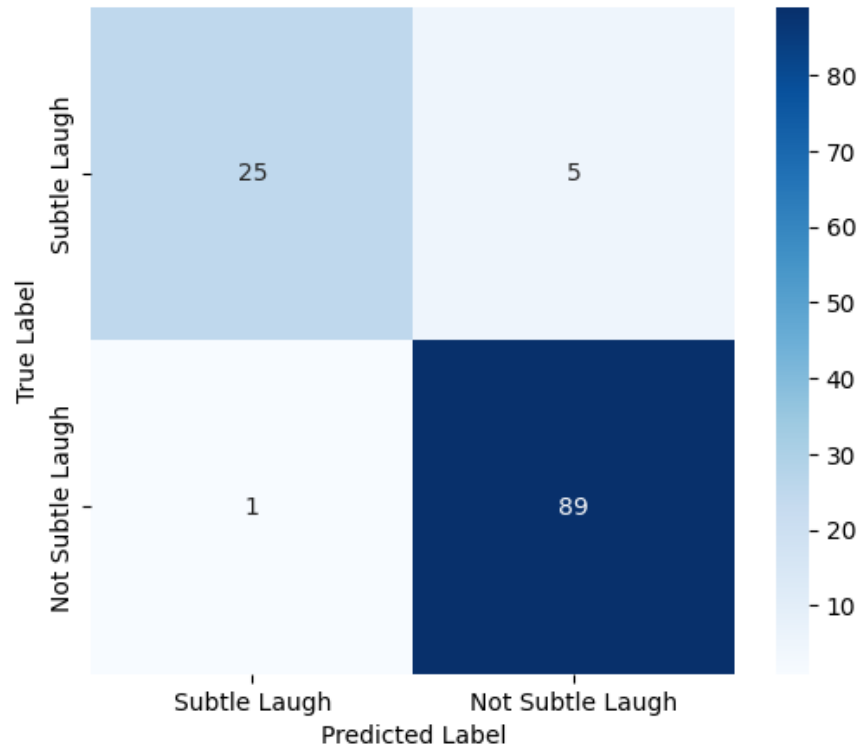


Figure 5.8: Confusion Matrix on the "Subtle Laugh" Expression Representation

### Surprised Representation

The system's ability to recognize and represent the Surprised expression through avatars in the virtual environment proved to be highly effective. It correctly identified 28 genuine instances of the surprised expression (True Positives), while only 1 non-surprised instance was mistakenly classified as surprised (False Positive). Additionally, it missed 2 actual occurrences of surprise (False Negatives), demonstrating a small margin for improvement in sensitivity. The system accurately recognized 89 cases where the surprised expression was not present (True Negatives), resulting in a strong accuracy score of 0.975. With a precision of 0.97, the system rarely misclassifies other expressions as surprise. The recall of 0.93 indicates that it effectively identifies most occurrences of the surprised expression, and the F1-score of 0.95 reflects a balanced and robust performance. Overall, these results demonstrate the system's strong capability to capture and convey the surprised expression accurately through avatar expressions, enhancing realism in virtual interactions.

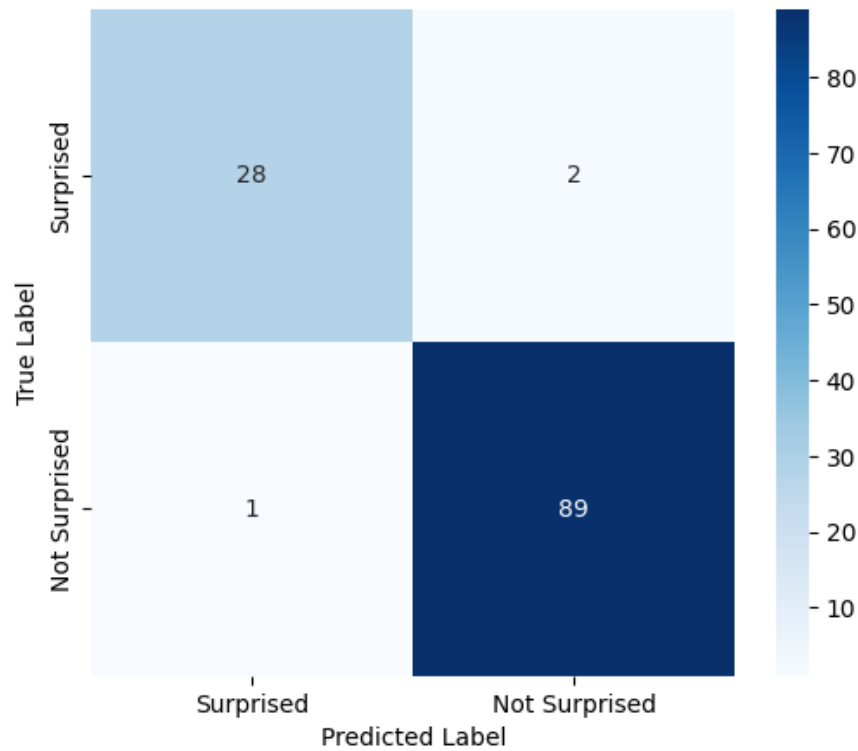


Figure 5.9: Confusion Matrix on the "Surprised" Expression Representation

### Neutral Representation

The system’s performance in recognizing and representing the Neutral expression through avatars in the virtual environment displayed a commendable ability to differentiate neutral from expressions. It accurately identified 29 neutral expressions (True Positives), with only 6 instances incorrectly classified as neutral when they were not (False Positives). However, it missed 1 neutral expression (False Negative), indicating a minor opportunity for improvement in detection. The system correctly rejected 84 non-neutral expressions (True Negatives), achieving an overall accuracy of 0.94, reflecting strong overall performance. The precision score of 0.83 shows that while the system is reasonably accurate when it classifies an expression as neutral, it occasionally mislabels other expressions as neutral. The recall of 0.97 indicates that the system is highly sensitive and identifies almost all neutral expressions, while the F1-score of 0.89 shows a solid balance between precision and recall. These results confirm that the system is proficient in recognizing and representing the neutral expression through avatars, contributing to a nuanced and realistic portrayal of user expressions in the virtual environment.

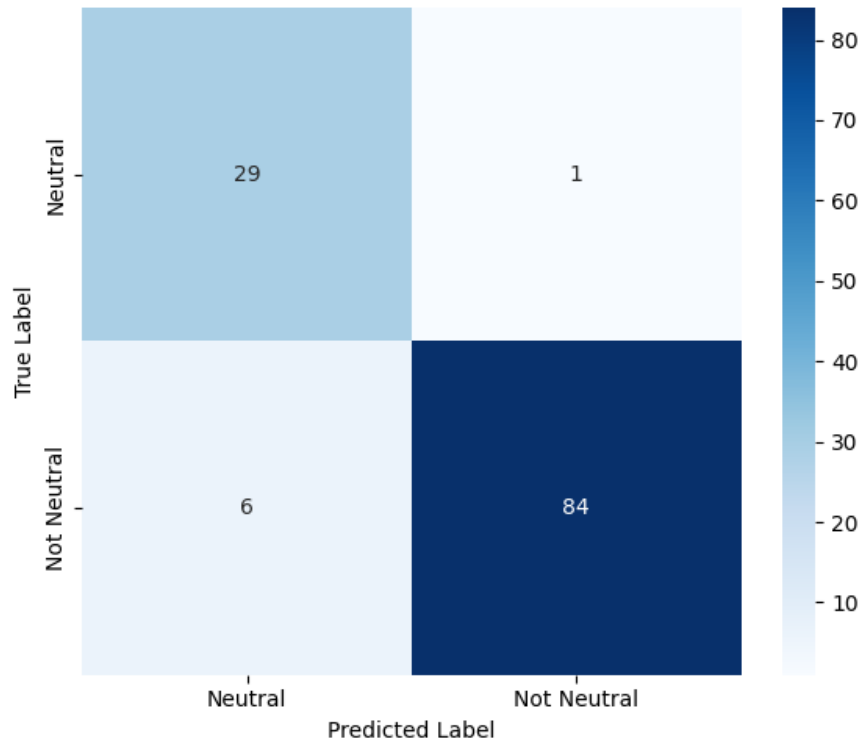


Figure 5.10: Confusion Matrix on the "Neutral" Expression Representation

In evaluating the system's performance across various expressions High Laugh, Subtle Laugh, Surprised, and Neutral the results highlight a generally strong ability to accurately recognize and represent these expressions in avatars within the virtual environment. The system demonstrated perfect performance for High Laugh, with an outstanding precision, recall, F1-score, and accuracy of 1.0. The Subtle Laugh expression showed strong performance, though slightly lower recall (0.83) and a slight drop in F1-score (0.89), indicating some room for improvement in capturing more subtle expressions. For Surprised, the system exhibited excellent precision (0.97) and recall (0.93), with a balanced F1-score of 0.95. Finally, the Neutral expression, while highly accurate (0.94), showed some room for improvement in precision (0.83) due to occasional misclassifications.

## 5.3 Visualizations

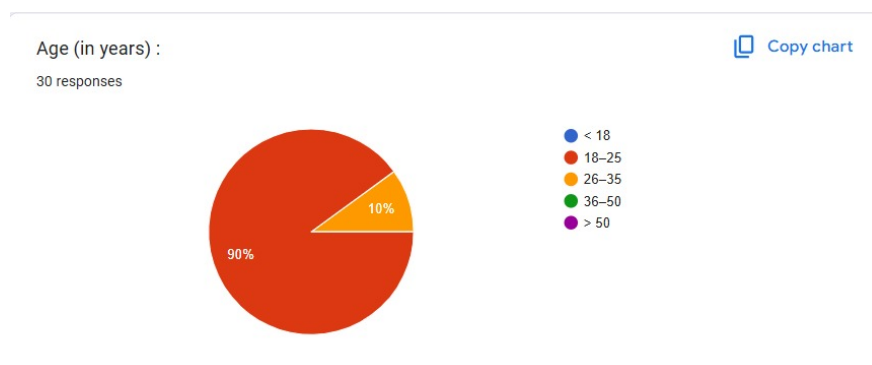


Figure 5.11: Distribution of Survey Respondents by Age Group

Figure 5.11 illustrates the age distribution of 30 survey respondents. The majority of participants (90%) fall within the 18–25 age group, indicating a predominantly young demographic. A smaller proportion, 10%, belongs to the 26–35 age range. No responses were recorded from individuals under 18, between 36–50, or over 50 years of age. This suggests that the findings of the study primarily reflect the perspectives and experiences of younger adults.

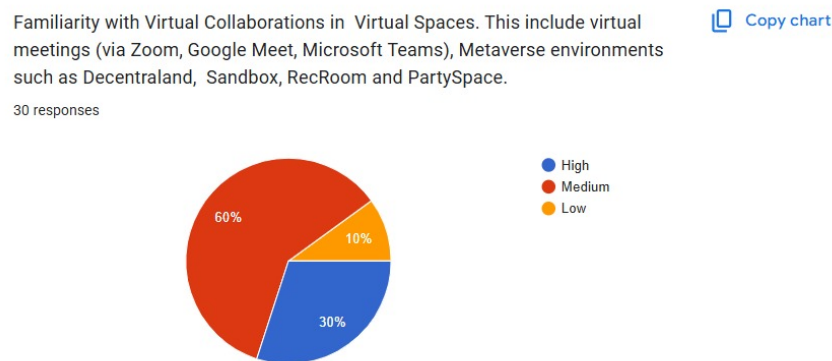


Figure 5.12: Familiarity Levels with Virtual Collaboration Tools Among Respondents

Figure 5.12 presents the level of familiarity among 30 respondents with virtual collaboration platforms, including virtual meeting tools (e.g., Zoom, Google Meet, Microsoft Teams) and metaverse environments (e.g., Decentraland, Sandbox, RecRoom, PartySpace). The majority of participants (60%) reported a medium level of familiarity, indicating a general awareness and moderate usage of such platforms. Meanwhile, 30% of respondents demonstrated a high level of familiarity, suggesting extensive experience and engagement. Only a small portion (10%) indicated low familiarity, reflecting limited exposure. These findings imply that most respondents possess a functional understanding of virtual collaboration technologies, which is beneficial for research focusing on digital interaction environments.

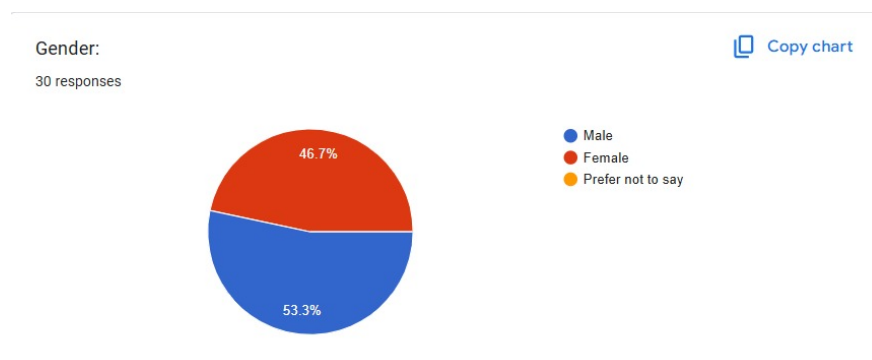


Figure 5.13: Gender Distribution of Survey Respondents

Figure 5.13 depicts the gender distribution among the 30 respondents who participated in the survey. A slight majority identified as male (53.3%), while 46.7% identified as female.

No respondents selected the option "Prefer not to say." The near-even distribution of genders suggests balanced representation, contributing to a more inclusive and diverse perspective within the research sample. This balance enhances the reliability of the study's findings across gender lines.

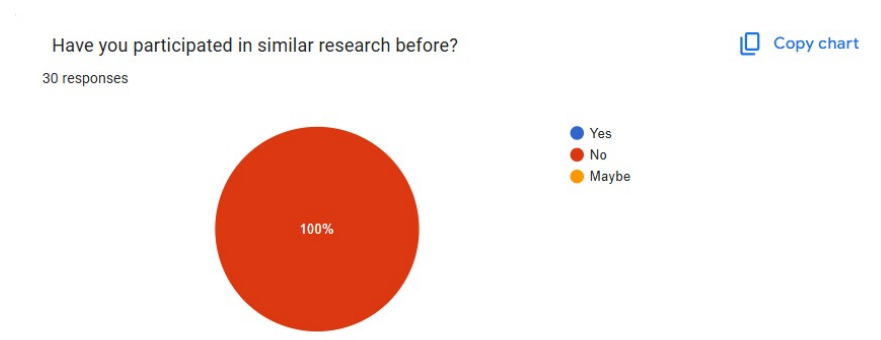


Figure 5.14: Participants' Prior Experience with Similar Research Studies

Figure 5.14 illustrates the responses to the question, "Have you participated in similar research before?", gathered from a total of 30 participants. The data reveals that 100% of the respondents indicated that they have not previously participated in similar research studies. This uniform response, represented entirely by the red section of the pie chart, suggests a lack of prior exposure or experience with related research among the sample population. Such a finding may imply that the insights collected through this study are likely to be shaped by fresh, unbiased perspectives, unaffected by prior involvement in similar research activities.

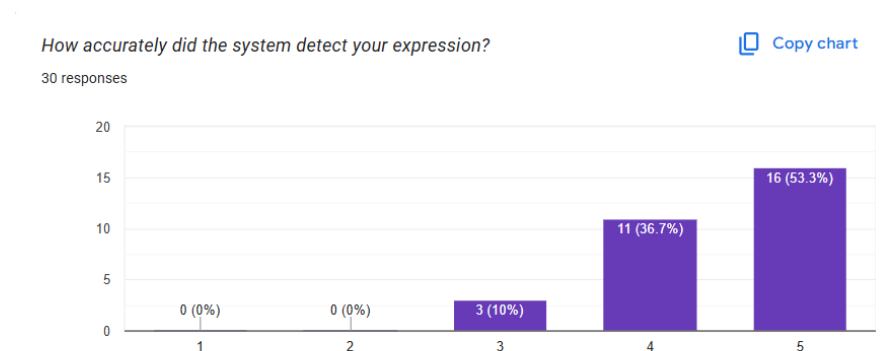


Figure 5.15: System accuracy in detecting "High Laugh" expressions.

Figure 5.15 shows the system's accuracy in detecting high laugh expressions received positive feedback from 30 respondents. A combined 90% of users rated the detection as good (4/5 - 53.3%) or excellent (5/5 - 36.7%). Only 10% gave a neutral rating of 3/5, with no negative ratings (1-2/5). This indicates strong performance in recognizing high-intensity laughter.

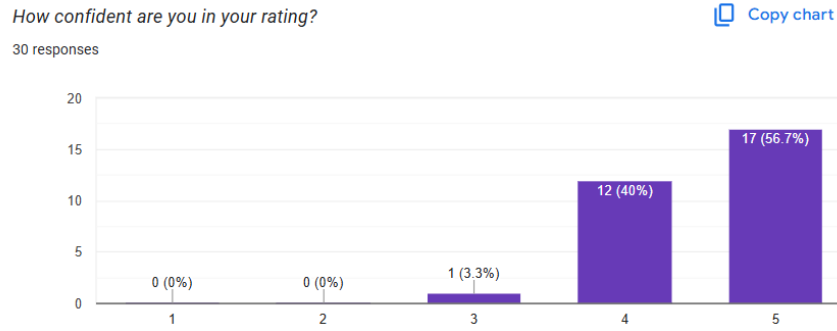


Figure 5.16: User confidence levels in "High Laugh" detection ratings

Figure 5.16 illustrates the levels of confidence participants had in the accuracy ratings they provided regarding the system's ability to detect facial expressions. The responses, collected from 30 participants using a 5-point Likert scale (1 = not confident at all, 5 = extremely confident), reveal a high degree of certainty among respondents. A significant majority—56.7% (17 participants)—expressed the highest level of confidence (rating of 5), while an additional 40% (12 participants) rated their confidence at level 4. Only one participant (3.3%) rated their confidence at level 3, and no participants selected the lowest confidence levels (1 or 2). These results indicate that the participants were generally self-assured in their evaluations of the system's performance.

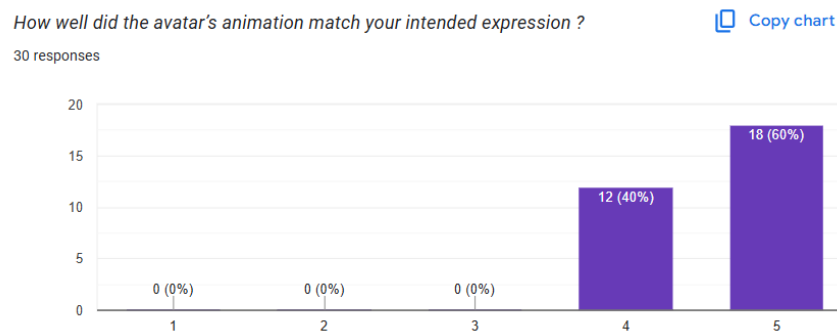


Figure 5.17: Avatar animation accuracy for "High Laugh" expressions

Figure 5.17 illustrates the results of a user study evaluating how accurately an avatar's animation represented the intended expression of the user. A total of 30 participants provided responses on a 5-point Likert scale, with 1 indicating "not at all accurate" and 5 indicating "perfectly accurate." The data show a highly positive response, with 60% of participants (18 respondents) rating the animation accuracy at the highest level (5), and an additional 40% (12 respondents) giving a rating of 4. Notably, no participants selected ratings of 1, 2, or 3, indicating a strong consensus regarding the avatar's effective conveyance of expression. These findings suggest that the avatar animation system performed reliably in matching users' intended expressions.



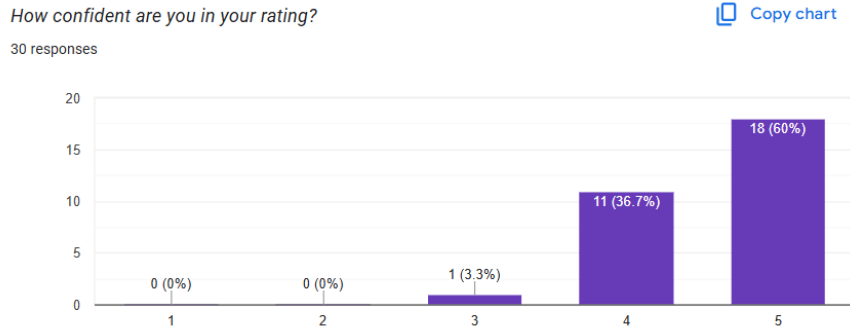


Figure 5.18: User confidence in "High Laugh" representation ratings.

Figure 5.18 Confidence in representation ratings was exceptionally high: 60% selected 5/5 and 36.7% chose 4/5. Only 3.3% rated confidence as 3/5, with no lower ratings. This demonstrates strong user certainty about the avatar's high laugh animation quality.

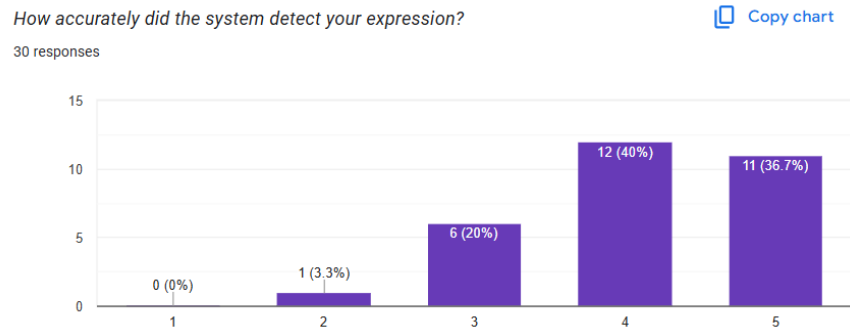


Figure 5.19: Accuracy of "Subtle Laugh" expression detection by the system

Figure 5.19 illustrates the system's accuracy in detecting subtle laugh expressions, based on 30 responses. The majority of users (76.7%) rated the detection positively, with 40% selecting 4/5 and 36.7% choosing 5/5. A smaller portion (20%) gave a moderate rating of 3/5, while only 3.3% rated it 1/5. This indicates generally good performance, though with room for improvement in consistency.

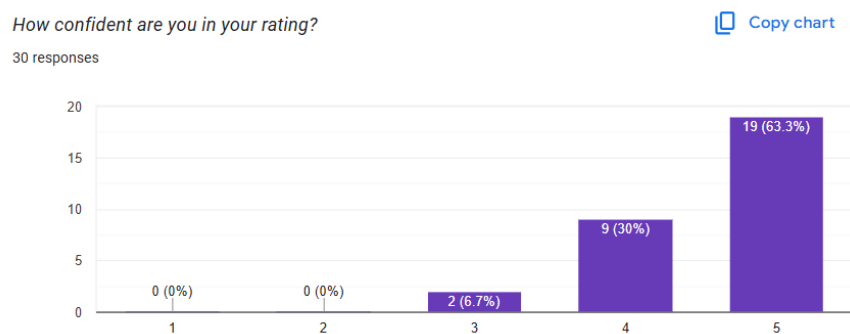


Figure 5.20: User confidence levels in their ratings of "Subtle Laugh" detection accuracy

In Figure 5.20 Respondents' confidence in their subtle laugh detection ratings was moderately high, with 63.3% selecting the highest confidence level (5/5) and 30% choosing 4/5. A small minority (6.7%) rated their confidence as 3/5, suggesting most users were fairly certain about their accuracy evaluations.

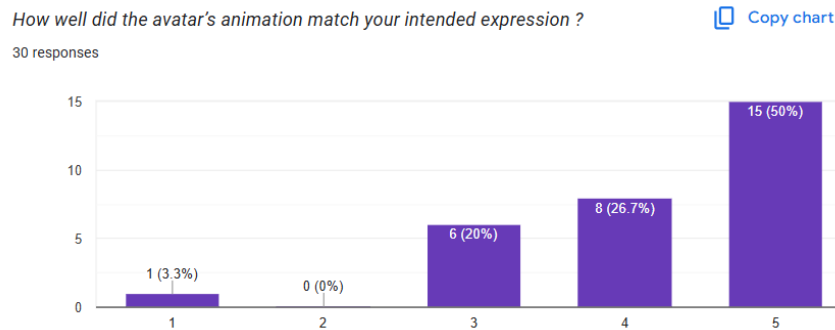


Figure 5.21: Alignment between avatar animations and user-intended "Subtle Laugh" expressions

In Figure 5.21 Users assessed how well the avatar's animations matched their intended subtle laugh expressions. Half (50%) gave the highest rating (5/5), while 26.7% selected 4/5. However, 20% rated it 3/5, and 3.3% chose 1/5, indicating noticeable variation in perception of the avatar's representation quality.

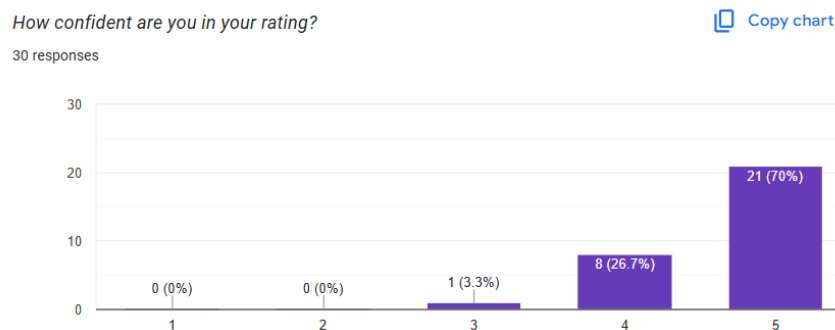


Figure 5.22: User confidence levels in their ratings of "Subtle Laugh" representation accuracy

Figure 5.22 shows Confidence in subtle laugh representation ratings was strong, with 70% selecting 5/5 and 26.7% choosing 4/5. Only 3.3% rated their confidence as 3/5, showing users were generally very certain about their evaluations of the avatar's performance.

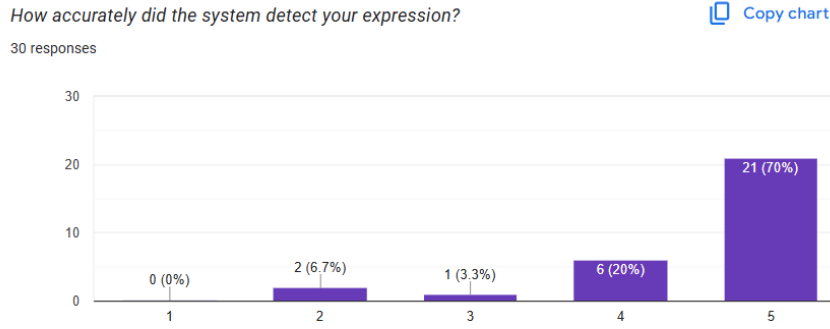


Figure 5.23: Accuracy of "Surprise" expression detection by the system

Figure 5.23 shows the system's performance in detecting surprise expressions, with 70% of users rating detection as highly accurate (5/5) and 20% giving a good rating (4/5). Only 10% reported moderate accuracy (3/5 or lower), indicating robust performance in recognizing surprise for most users.

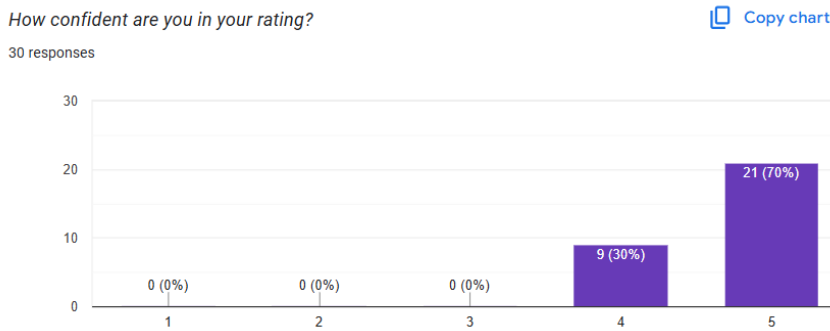


Figure 5.24: User confidence levels in their ratings of "Surprise" detection accuracy

Figure 5.24 shows high confidence in surprise detection ratings, with 70% selecting the highest confidence level (5/5) and 30% choosing 4/5. The absence of ratings below 4 demonstrates strong user assurance in their evaluations of the system's surprise detection capabilities.

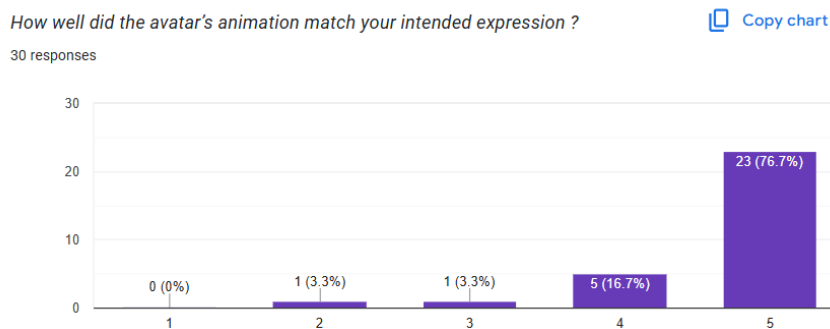


Figure 5.25: Alignment between avatar animations and user-intended "Surprise" expressions

Above 5.25 shows that the majority (76.7%) gave perfect scores (5/5) for how well the avatar matched their surprise expressions, while 16.7% rated it 4/5. A small minority (6.6% combined) reported lower matches, suggesting only minor discrepancies for a few users.

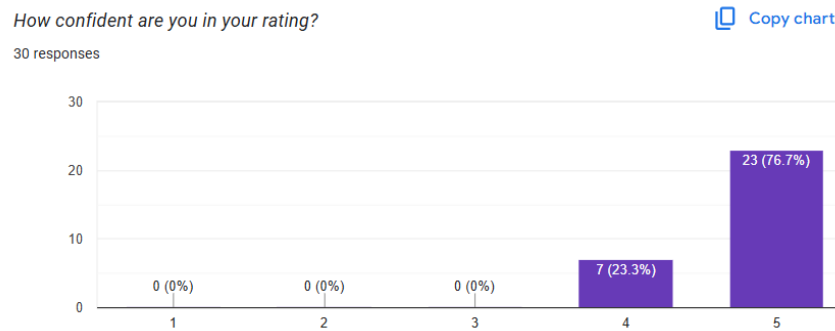


Figure 5.26: User confidence levels in their ratings of "Surprise" representation accuracy

Figure 5.26 mirrors detection confidence, with 76.7% selecting 5/5 and 23.3% choosing 4/5. The complete absence of lower confidence ratings reinforces reliable user judgments about the avatar's surprise animations.

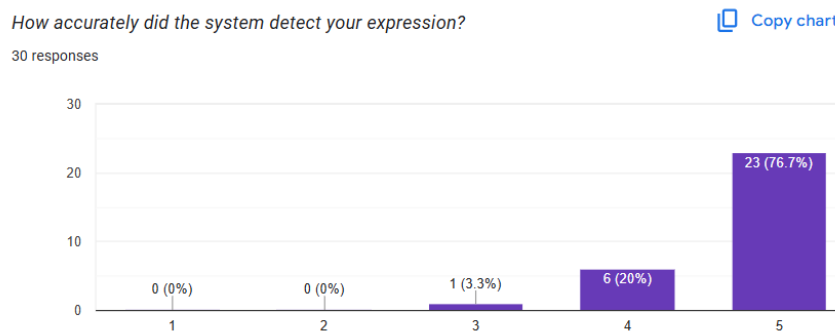


Figure 5.27: Accuracy of "Neutral" expression detection by the system.

Figure 5.27 shows that 76.7% of users rated the system's neutral expression detection as highly accurate (5/5), while 20% gave it a good rating (4/5). Only 3.3% reported moderate accuracy (3/5), with no lower ratings, indicating excellent performance in recognizing neutral expressions.

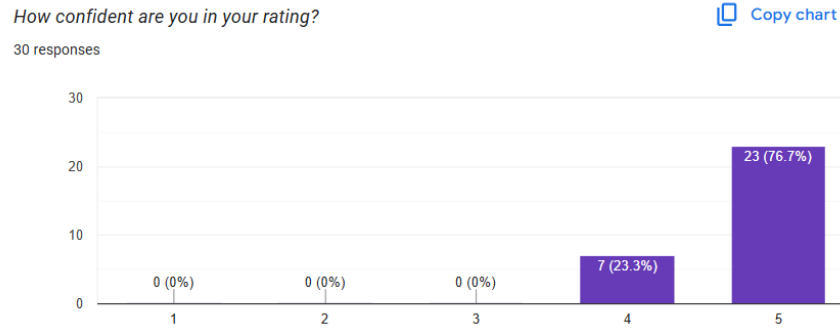


Figure 5.28: User confidence levels in their ratings of "Neutral" expression detection accuracy.

Figure 5.28 demonstrates strong user confidence, with 76.7% selecting the highest confidence level (5/5) and 23.3% choosing 4/5. The complete absence of ratings below 4 shows users were very certain about their evaluations of the system's neutral expression detection.

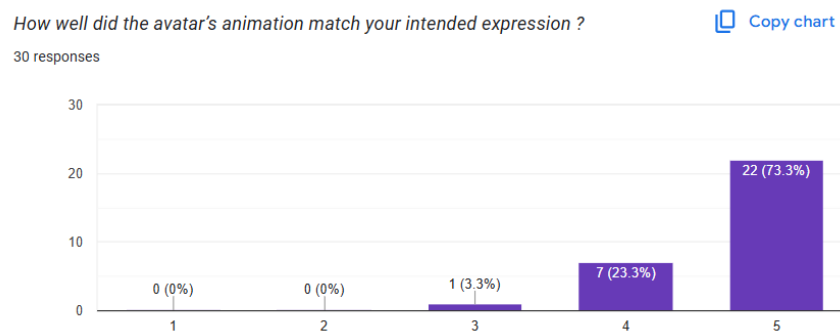


Figure 5.29: Alignment between avatar animations and user-intended "Neutral" expressions.

Figure 5.29 reveals that 73.3% of users gave perfect scores (5/5) for how well the avatar matched their neutral expressions, while 23.3% rated it 4/5. Only 3.3% reported lower matches, suggesting nearly all users found the animations accurate.

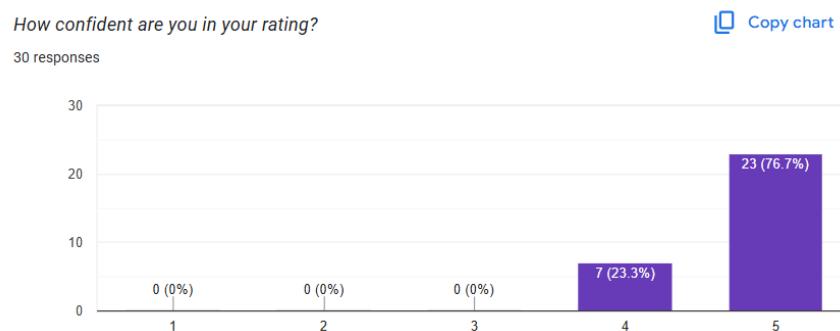


Figure 5.30: User confidence levels in their ratings of "Neutral" expression representation accuracy.

Figure 5.30 shows identical confidence levels to detection ratings, with 76.7% selecting 5/5 and 23.3% choosing 4/5. The consistent high confidence across both detection and representation ratings indicates reliable user assessments of the system's neutral expression handling.

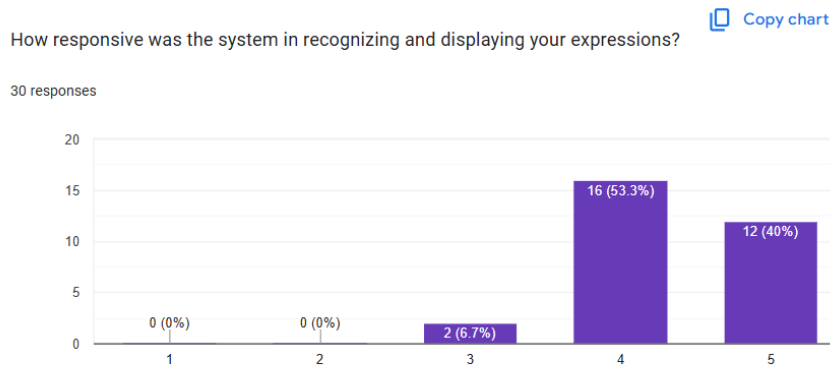


Figure 5.31: Participant Ratings on the responsiveness of recognizing and displaying expressions

Figure 5.31 reflects feedback from 30 respondents on the system's responsiveness in recognizing and displaying user expressions, rated on a scale from 1 (not responsive) to 5 (highly responsive). The majority of users (93.3%) reported positive experiences, with 53.3% rating the system a 5 (highly responsive) and 40% rating it a 4. A small minority (6.7%) gave it a neutral rating of 3, while no respondents selected ratings of 1 or 2. This indicates strong performance in responsiveness, with nearly all users finding the system effective at recognizing and displaying expressions.

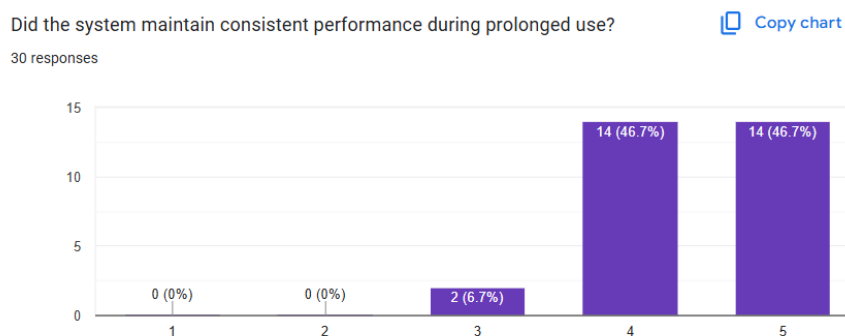


Figure 5.32: Participant Ratings on the system maintaining consistent performance.

Figure 5.32 presents participant feedback regarding the consistency of system performance over extended periods of use. The responses were collected from 30 participants using a 5-point Likert scale, where 1 indicates "Unusable due to delays" and 5 indicates "Yes, performance remained consistent". The data reveal that 93.4% of participants rated the system's performance as either 4 (46.7%) or 5 (46.7%), demonstrating a high level of perceived consistency. Only 6.7% (2 participants) rated the system at level 3, and none

reported ratings of 1 or 2. These results strongly suggest that the system maintained reliable and stable performance throughout prolonged usage, highlighting its robustness and potential suitability for sustained interactions

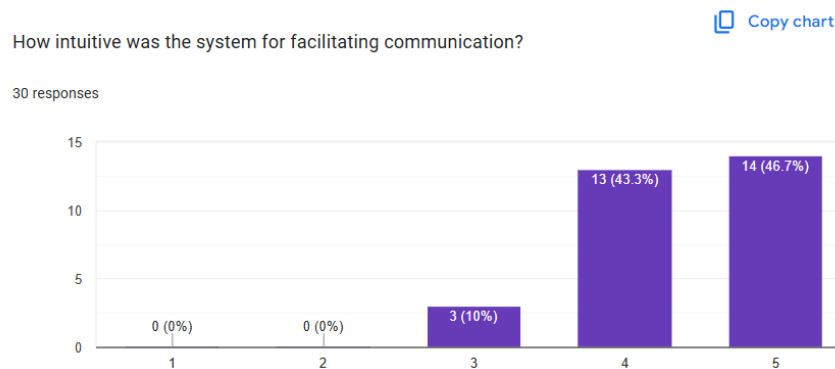


Figure 5.33: Participant Ratings on the intuitive communication.

Figure 5.33 illustrates user perceptions of a system's intuitiveness for facilitating communication, based on 30 responses rated on a scale from 1 (not intuitive) to 5 (very intuitive). The results show overwhelmingly positive feedback, with 90% of respondents rating the system a 4 (43.3%) or 5 (46.7%), indicating it was highly intuitive for communication purposes. A small minority (10%) gave it a neutral rating of 3, while no respondents selected ratings of 1, or 2. This suggests the system was well-received, with nearly all users finding it intuitive and effective for communication.

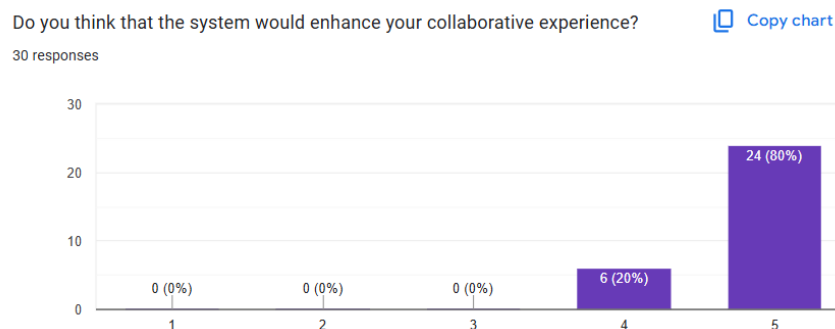


Figure 5.34: Participant Ratings on the System's Potential to Enhance Collaborative Experience.

The figure 5.34 depicts user responses concerning the system's perceived ability to enhance collaborative experiences. Based on feedback from 30 participants using a 5-point Likert scale—where 1 indicates "Strongly Disagree" and 5 indicates "Strongly Agree" the results show overwhelmingly positive sentiment. A substantial 80% of respondents (24 participants) rated the system at the highest level (5), while the remaining 20% (6 participants) selected a rating of 4. No participants rated the system below 4, indicating unanimous agreement on the system's value in supporting and enriching collaborative

interactions. These findings suggest that the system is highly effective in fostering collaborative engagement and could serve as a valuable tool in cooperative environments.

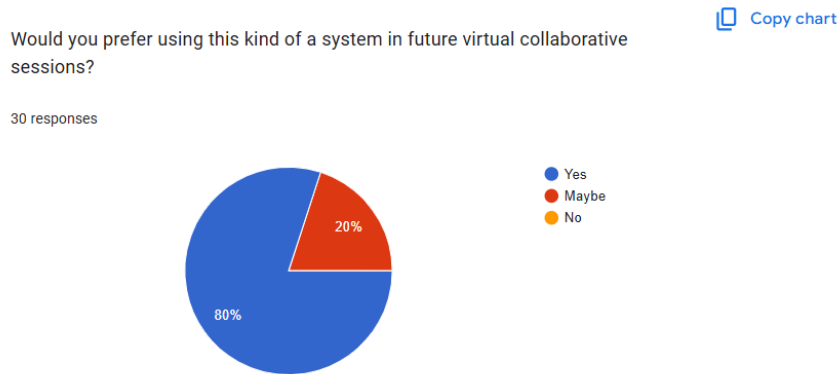


Figure 5.35: User Preference for Future Adoption of the System in Virtual Collaboration.

The figure 5.35 shows responses from 30 participants regarding their willingness to use the system in future virtual collaborative sessions. A significant majority (80%) answered "Yes," indicating strong approval and a high likelihood of continued usage. Meanwhile, 20% responded with "Maybe," suggesting some uncertainty or conditional acceptance, while no respondents selected "No." This reflects a generally positive reception, with most users inclined to adopt the system again for collaborative work.

## 5.4 Comparison with Existing Work

This section contextualizes the developed system, VirExp, within the broader research landscape by comparing its approach and performance to existing work in terms of: (1) Similar expression detection systems' reported accuracy metrics, (2) Avatar animation techniques in current virtual collaboration platforms, and (3) Established benchmarks for real-time expression recognition systems in virtual environments.

### 5.4.1 Accuracy Metrics Comparison

Similar expression detection systems have reported varying levels of accuracy. [7] utilized an ensemble decision tree to identify emotions based on body gestures, achieving an overall accuracy of 90.83%. [20]'s work, focusing on emotion recognition solely from upper-body movements, reported an accuracy of 76.8%. [8] suggested a solution that reached approximately 90% accuracy in recognizing movements. [9]'s work, while achieving a high classification accuracy of 97%, was limited to considering only upper-body movements.

However, considering the results of this study's pipeline, VirExp performs comparably to or better than some existing systems that focus on subsets of expressions or different modalities, while integrating both facial and upper-body cues for real-time recognition.



### 5.4.2 Avatar Animation Technique Comparison

Current virtual collaboration platforms and existing research employ various techniques for avatar animation to represent user expressions. Many platforms rely on manual input or predefined animations, which limits the spontaneity and authenticity of interactions. Platforms like Party.Space, Decentraland, and Rec Room utilize avatars in virtual environments for social and collaborative interactions, but often depend on users manually selecting animations.

The VirExp pipeline focuses on automating the process of translating real-world physical expressions into dynamic, real-time avatar animations. This utilizes VRoid characters and Mixamo animations within a Unity multiplayer framework to represent detected expressions. Exaggerated animations were employed to enhance expression recognizability in the virtual environment. The system captures user expressions via webcam, processes them, and transmits data to Unity to trigger corresponding avatar animations, aiming for seamless and near real-time representation. This approach moves beyond static or manually triggered animations to create more immersive and expressively rich interactions.

### 5.4.3 Established Benchmarks for Real-Time Expression Recognition Systems in Virtual Environments

Established benchmarks in real-time expression recognition often involve the use of specific datasets and evaluation metrics. Datasets like GEMEP and FER-2013 have been utilized in related studies. Performance is typically evaluated using metrics such as precision, recall, F1-score, and accuracy, particularly in real-time scenarios. Also, some studies like [7] uses Kinect devices to capture skeletal points of the user’s body. Therefore, the ability to perform recognition in near real-time without requiring sophisticated or expensive hardware is a significant consideration.

VirExp utilizes a dataset that is created from scratch, and as mentioned in the Evaluation and Results sections, metrics like precision, recall, F1-score, etc. were used to evaluate the performance. However, none of the sophisticated infrastructure was used for the VirExp pipeline, and only a simple web cam/inbuilt cam was used.

# Chapter 6

## Discussion and Recommendations

### 6.1 Discussion

The results of this research support all the stated objectives and the three research questions that were planned to address. The aim in this section is to demonstrate how the developed system addresses the identified problem of limited physical expressiveness in virtual environments and enhances collaborative interactions.

Based on the research conducted, specific facial expressions and body gestures were identified for the target expressions of High Laugh, Subtle Laugh, Surprise, and Neutral. For instance, High Laugh was associated with mostly closed or less open eyes, raised cheeks, wrinkles around the eye corners, and a widely opened mouth, accompanied by extended body, upper body movements, shoulder trembling, and forward head movements. Surprise was characterized by raised eyebrows, wide-open eyes, and a widely open mouth or dropped jaw, with quick backward movements or sudden head/hand gestures. The MediaPipe Holistic Library was utilized to capture the corresponding skeletal point sequences for these expressions, providing 543 landmarks covering pose, face, and hand landmarks. This directly addresses the first research question by identifying the physical cues and their associated skeletal data for distinct expressions. Identifying these specific facial expressions and body gestures related to different expressions and ways of capturing skeletal point sequences for each movement addresses RQ 1, which is "What specific facial expressions and body gestures convey distinct expressions, and what are the corresponding sequences of skeletal points associated with these gestures?". Furthermore, the research fulfills three key objectives. First, the objective of "Understanding which expressions are displayed and identified in a collaborative environment" is achieved by defining and classifying High Laugh, Subtle Laugh, Surprise, and Neutral as recognizable expressions. Second, the objective of "Understanding the connection between facial expressions, body gestures, and skeletal point sequences in conveying expressions" is met by establishing concrete relationships between visual cues (e.g., raised cheeks, wide-open eyes) and their corresponding skeletal landmarks. Third, the objective of "Capturing and analyzing data on skeletal point sequences in conjunction with facial expressions in conveying expressions" is accomplished through the structured collection and processing of landmark data using MediaPipe Holistic.

The study developed an automated pipeline, VirExp, using an LSTM model to identify predefined facial expressions and body gesture patterns from real-time skeletal point sequences. The evaluation demonstrated how well the system is able to find all relevant instances and avoid irrelevant instances with a high F1 score of 91.67%. Specifically, the LSTM model showed strong performance in classifying High Laugh (F1-score: 94.91%), Subtle Laugh (F1-score: 91.23%), Surprise (F1-score: 91.53%), and Neutral (F1-score: 89.23%). The integration with Unity via FastAPI allowed for the near real-time transmission of identified expression labels, triggering corresponding avatar animations. Also, the time taken for expression detection was tested, where the LSTM-based system and Transformer-based system were performing within near real time, while the DTW-based system took between 45 seconds to 50 seconds to identify expressions. Therefore, out of these three solutions, the ones with the lowest detection time, which are the LSTM and Transformer-based models, were selected as potential pattern recognition technologies to satisfy the near-real-time aspect of the system. Next, resource usage was considered. Out of LSTM and Transformer-based models, LSTM was more lightweight than the Transformer and was able to run in a normal day-to-day setup. For that reason, the LSTM-based model was recognized as a more suitable pattern recognition technique for the pipeline that was implemented. This study also evaluated expression representation accuracy as well, where it was able to get a high F1 score of 93.33% indicating that users were able to represent their expression as expected to the other party using VirExp. These facts confirm that RQ 2, which is "In a real-time skeletal point sequence, how can we identify predefined facial expressions and body gesture patterns within near real-time and express them?" is met. The study also successfully achieved all related research objectives. First, it fulfilled the objective of "Understanding on how a virtual avatar expresses the identified expressions" by creating a system that accurately translates detected expressions into appropriate avatar animations in Unity. Second, it accomplished the goal of "Building and evaluating an automated pipeline that uses these patterns for expression recognition and representation in a virtual environment" through the development and thorough testing of VirExp. Finally, the research met the objective of "Designing a method to represent expressions with varying intensities on an avatar in the virtual space" by implementing dynamic animations that adjust based on identifying and representing high laugh which is the high intensity expression in contrast with the subtle laugh which represents the low intensity expression.

Based on the evaluation and results, the suggested solution demonstrates a significant capacity to perform effectively and help users facilitate collaborative interactions in the virtual space. This is evidenced by both the system's technical performance in expression recognition and representation, as verified and validated in the above paragraphs, and the positive feedback received from users regarding their experience. User feedback supports the conclusion that the system facilitates collaborative interactions. Participants generally reported positive experiences regarding the system's responsiveness in recognizing and displaying expressions, with a large majority rating it as highly responsive (93.3% of the

participants giving scores of 4 and 5, in a scale of 1 to 5, 1 being less responsive and 5 being highly responsive). Also, the opinion regarding the system maintaining consistent performance during prolonged use was positive, with 93.4% of the participants stating the system was consistently performing, and the feedback on intuitive communication was overwhelmingly positive. Most users found the system intuitive and effective for communication, where 90% of users rated scores of 4 and 5 in a scale of 1 to 5, 1 being less intuitive and 5 being highly intuitive. A significant majority of participants (80% of the 30 participants) also expressed willingness to adopt such a system in future virtual collaborative sessions, underscoring its perceived value in enhancing virtual interactions. With these results RQ 3, which is "To what extent will the suggested solution perform and help users facilitate collaborative interactions in the virtual space?" is being addressed. Furthermore, these findings successfully meet the final research objective to "Evaluate the effectiveness and performance of the system in facilitating user collaboration." The combination of quantitative performance data and qualitative user satisfaction metrics provides robust validation that the system achieves its intended purpose of improving virtual collaborative experiences. The strong user acceptance (80% adoption willingness) particularly underscores the practical value and usability of the solution in actual collaborative settings.

## 6.2 Conclusion

This research successfully addressed the identified problem of limited physical expressiveness in virtual environments by developing and evaluating an automated process for real-time facial expression and upper body gesture recognition and representation. The study's objectives, aimed at enhancing physical expressions and communication within collaborative virtual spaces, were met through a rigorous design science research methodology.

The research successfully answered the proposed questions:

1. Specific facial expressions and body gestures conveying distinct expressions (High Laugh, Subtle Laugh, Surprise, and Neutral) and their corresponding skeletal point sequences were identified using the MediaPipe Holistic Library.
2. The developed process, Virexp, utilizing an LSTM-based pattern recognition model and integrated with Unity via FastAPI, demonstrated the ability to identify and express these patterns within near real-time, with high overall accuracy (91.67%) and efficient processing times.
3. The suggested solution's performance and its effectiveness in facilitating collaborative interactions were evaluated through comprehensive data analysis and user studies. The system achieved high F1 scores for both expression recognition (overall 91.67%) and representation (overall 93.33%), indicating its reliability in conveying intended expressions. User feedback highlighted enhanced collaborative experience, intuitive design, and a willingness to adopt the system for future virtual interactions.

Overall, this research contributes a valuable, accessible process for improving non-verbal communication in virtual settings. By automatically translating real-world physical expressions into dynamic avatar animations, the system enhances the sense of presence and connection, moving virtual collaboration closer to the richness of face-to-face interaction. While areas for future refinement exist, such as expanding the dataset and incorporating a wider range of expressions, this study provides a strong foundation for more expressive and engaging virtual environments.

## 6.3 Recommendations and Future Work

While the research successfully developed and evaluated the process created for real-time expression recognition and representation, there are areas for further improvement and future work.

To improve the robustness and applicability of the process, it is essential to expand the dataset to include greater demographic diversity. This means incorporating a wider range of age groups, ensuring gender inclusion, and representing varied cultural backgrounds to better assess how well the model generalizes between different populations. Furthermore, integrating professional actor performances could improve the quality of training data by providing controlled, high-fidelity expressions that are more consistent and precise. Increasing the sample size to over 100 participants would also strengthen statistical power and help reduce potential biases in the data. Future efforts could involve collaborating with multicultural research teams to gather region-specific expression data.

Further refinement of the model should focus on improving expression granularity by expanding the range of expressions it can recognize. This includes incorporating compound expressions, such as "happy surprise," as well as negative expressions like anger and sadness, to broaden the system's applicability. Testing adaptive intensity thresholds could also help account for individual differences in expressiveness, ensuring the model remains accurate across users with varying levels of expressions.

Enhancing user experience requires deeper qualitative analysis through structured open-ended prompts that capture nuanced feedback, such as instances where the avatar's expressions felt mismatched. Longitudinal studies would also be valuable to assess the long-term psychological effects of prolonged system use, including potential expressive fatigue or shifts in trust toward avatar representations. Future research could explore cultural adaptations, such as modifying eyebrow movements for expressions like "surprise" to better align with East Asian versus Western norms. Additionally, evaluating the system's performance in non-collaborative contexts, such as education or mental health therapy, could reveal new use cases and areas for improvement.

On the technical side, exploring multimodal fusion such as combining facial, upper body, and vocal cues could enhance expressiveness while maintaining hardware simplicity, with future implementations potentially incorporating adaptive personalization to better align with individual expressiveness. Developing plug-in standards for popular game engines

like Unity and Unreal would further facilitate broader adoption, making the system more accessible to developers. Beyond these enhancements, future research could expand the range of detectable expressions, moving beyond basic expressions to include more nuanced and compound expressional states. Refining the models to better distinguish subtle differences, such as between a Subtle Laugh and a Neutral expression, would significantly improve reliability, while exploring hybrid pattern recognition approaches could unlock further performance gains. Additionally, integrating deeper contextual understanding would enable more sophisticated responses in complex social scenarios. Long-term studies assessing psychological impacts and evaluating the system's applicability across diverse virtual interactions from education to mental health therapy will be essential for refining its real-world utility. This scalable approach ensures continuous improvement in accuracy, cultural adaptability, and practical deployment, supporting the system's evolution across an expanding range of domains.

# References

- [1] G. Fernandes, J. Barbosa, E. B. Pinto, M. Araújo, and R. J. Machado, “Applying a method for measuring the performance of university-industry r&d collaborations: case study analysis,” *Procedia Computer Science*, vol. 164, pp. 424–432, 2019.
- [2] T. Greven, “The influence of non-verbal behaviour on meeting effectiveness and pro-active behaviour: A video observational study,” B.S. thesis, University of Twente, 2017.
- [3] B. Deng, “The impact of the metaverse on the future development of enterprises,” *BCP Business & Management*, 2023.
- [4] E. Spadoni, M. Carulli, M. Mengoni, M. Luciani, and M. Bordegoni, “Empowering virtual humans’ emotional expression in the metaverse,” in *International Conference on Human-Computer Interaction*, pp. 133–143, Springer, 2023.
- [5] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 59–66, IEEE, 2018.
- [6] R. Santhoshkumar and M. K. Geetha, “Deep learning approach for emotion recognition from human body movements with feedforward deep convolution neural networks,” *Procedia Computer Science*, vol. 152, pp. 158–165, 2019.
- [7] S. Saha, S. Datta, A. Konar, and R. Janarthanan, “A study on emotion recognition from body gestures using kinect sensor,” in *2014 international conference on communication and signal processing*, pp. 056–060, IEEE, 2014.
- [8] T. Chaves, L. Figueiredo, A. Da Gama, C. de Araujo, and V. Teichrieb, “Human body motion and gestures recognition based on checkpoints,” in *2012 14th Symposium on Virtual and Augmented Reality*, pp. 271–278, IEEE, 2012.
- [9] Y. Xiao, J. Yuan, and D. Thalmann, “Human-virtual human interaction by upper body gesture understanding,” in *Proceedings of the 19th ACM symposium on virtual reality software and technology*, pp. 133–142, 2013.
- [10] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, “Real-time automatic emotion recognition from body gestures,” *arXiv preprint arXiv:1402.5047*, 2014.

- [11] S. Scotti, M. Mauri, R. Barbieri, B. Jawad, S. Cerutti, L. Mainardi, E. N. Brown, and M. A. Villamira, “Automatic quantitative evaluation of emotions in e-learning applications,” in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1359–1362, IEEE, 2006.
- [12] L. E. Nacke, M. Kalyn, C. Lough, and R. L. Mandryk, “Biofeedback game design: using direct and indirect physiological control to enhance game interaction,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 103–112, 2011.
- [13] I. d. L. Ortiz-Vigon Uriarte, B. Garcia-Zapirain, and Y. Garcia-Chimeno, “Game design to measure reflexes and attention based on biofeedback multi-sensor interaction,” *Sensors*, vol. 15, no. 3, pp. 6520–6548, 2015.
- [14] C. Kerdvibulvech, “A digital human emotion modeling application using metaverse technology in the post-covid-19 era,” in *International Conference on Human-Computer Interaction*, pp. 480–489, Springer, 2023.
- [15] V. Gentile, F. Milazzo, S. Sorce, A. Gentile, A. Augello, and G. Pilato, “Body gestures and spoken sentences: a novel approach for revealing user’s emotions,” in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pp. 69–72, IEEE, 2017.
- [16] F. Ahmed, A. H. Bari, and M. L. Gavrilova, “Emotion recognition from body movement,” *IEEE Access*, vol. 8, pp. 11761–11781, 2019.
- [17] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, “Technique for automatic emotion recognition by body gesture analysis,” in *2008 IEEE Computer society conference on computer vision and pattern recognition workshops*, pp. 1–6, IEEE, 2008.
- [18] V. Sekar and A. Jawaharlalnehru, “Semantic-based visual emotion recognition in videos-a transfer learning approach,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3674–3683, 2022.
- [19] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, “Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings,” in *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 99–106, 2009.
- [20] C. M. A. Ilyas, R. Nunes, K. Nasrollahi, M. Rehm, and T. B. Moeslund, “Deep emotion recognition through upper body movements and facial expression.,” in *VISIGRAPP (5: VISAPP)*, pp. 669–679, 2021.
- [21] P. Metri, J. Ghorpade, and A. Butalia, “Facial emotion recognition using context based multimodal approach,” 2011.



- [22] H. Gunes and M. Piccardi, “Bi-modal emotion recognition from expressive face and body gestures,” *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.
- [23] I. Ralev and G. Krastev, “Application of opencv in serious games,” in *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 484–487, IEEE, 2022.
- [24] A. Jungherr and D. B. Schlarb, “The extended reach of game engine companies: How companies like epic games and unity technologies provide platforms for extended reality applications and the metaverse,” *Social Media+ Society*, vol. 8, no. 2, p. 20563051221107641, 2022.
- [25] G. Dumbravă and A. Koronka, ““actions speak louder than words”-body language in business communication,” *OF THE UNIVERSITY OF PETROȘANI ECONOMICS*, vol. 9, no. 3, pp. 249–254, 2009.
- [26] K. Wang, C. Ishi, and R. Hayashi, “Preliminary analysis of facial expressions and body movements of four types of laughter,” in *Proc. LW 2024*, pp. 21–23, 2024.
- [27] Y. Tang, Q. Li, and X. Chen, “Fastapi: Benchmarking lightweight python web frameworks,” *Journal of Systems Software*, vol. 183, p. 111104, 2021.
- [28] K. Zhang, L. Wang, and M. Zheng, “Frame-skipping for real-time action recognition,” in *CVPR Workshops*, pp. 12–21, 2020.
- [29] P. Ekman and W. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, 1978.

# Appendices

# VirExp Survey: Enhancing Emotional Expressiveness in Virtual Collaborations

We are a group of final-year undergraduates from the University of Colombo School of Computing, and we are excited to invite you to participate in our research survey! Our research, "

VirExp : Automated Emotion and Gestures for Virtual Collaborative Environments," aims to enhance emotional expressiveness in virtual collaborations by automatically recognizing and conveying emotions through facial expressions and body gestures. We want to understand better your current experience related to virtual collaborations and needs to create a more engaging and effective virtual collaboration solution.

Your participation in this survey will help us gather valuable insights into the requirements and experiences of users like you. The survey is quick, and your responses will be used solely for academic purposes.

We would greatly appreciate your time and effort in completing this survey. Your feedback is crucial in helping us develop innovative solutions for virtual collaborations. Thank you in advance for your precious time and effort!

1. Email \*

---

2. How often do you participate in virtual meetings / collaborations ?

*Mark only one oval.*

1   2   3   4   5

---

Never ☐ ☐ ☐ ☐ ☐ Often

---

## 3. Do you feel that virtual collaborations lack emotional expressiveness?

*Mark only one oval.*

	1	2	3	4	5	
No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Yes

## 4. How important is non-verbal communication (e.g., facial expressions, gestures) in your virtual interactions?

*Mark only one oval.*

	1	2	3	4	5	
Not	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Important

## 5. Have you experienced misunderstandings due to lack of emotional cues in virtual collaborations?

*Mark only one oval.*

☐ Yes

☐ No

## 6. How often do you switch on video during virtual collaborations?

*Mark only one oval.*

	1	2	3	4	5	
Never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Always

7. How comfortable are you with switching on video during virtual collaborations?

*Mark only one oval.*

1   2   3   4   5

Not ☐ ☐ ☐ ☐ ☐ Very comfortable

8. Would you find it useful if a system could automatically display your emotions via an avatar in virtual collaborations?

*Mark only one oval.*

☐ Yes

☐ No

☐ Maybe

9. Would you prefer a system that recognizes and displays emotions automatically over manual options like selecting emojis?

*Mark only one oval.*

☐ Yes

☐ No

☐ Maybe

10. Do you think automated emotion recognition could improve your engagement in virtual collaborations?

*Mark only one oval.*

☐ Yes

☐ No

☐ Maybe

# Participant Information and Consent Form

**Study Title:**

VirExp: Automated Emotion and Gestures for Virtual Collaborative Environment

**Researcher Contact Information:**

**Name:** Masha Pupulewatte, Chris M. Perera, Rashmina Senadheera

**Email:** [2020is077@stu.ucsc.cmb.ac.lk](mailto:2020is077@stu.ucsc.cmb.ac.lk), [2020is075@stu.ucsc.cmb.ac.lk](mailto:2020is075@stu.ucsc.cmb.ac.lk), [2020is@096stu.ucsc.cmb.ac.lk](mailto:2020is@096stu.ucsc.cmb.ac.lk)

**Phone:** 076-3397994, 077-0562741, 077-4550849

**Study Overview:**

You are invited to participate in a research study aimed at identifying and representing emotions through skeletal point sequences of specific facial expressions and gestures to enhance emotional expressions in virtual environments.

**What will you be asked to do?**

- Perform four gestures: **High laugh, subtle laugh, confused, and neutral.**
- Each gesture will be performed 30 times under supervised conditions.
- No video recordings of your gestures will be stored; only skeletal point sequences will be collected and stored as/compared with numpy arrays.

**How will your data be used?**

- The data will be used **exclusively for research purposes.**
- No data will be shared with external parties.

**Participation is voluntary.**

You may withdraw at any time without any consequences.

\* Indicates required question

## Participant Information

1. Full name with initials

Eg : A.B.C. Perera

---

2. Age

---

## 3. Gender

*Mark only one oval.*

☐ Male

☐ Female

☐ Other: \_\_\_\_\_

## 4. Email Address (Optional)

\_\_\_\_\_

## 5. How often are you interacting in virtual collaborations?

*Mark only one oval.*

☐ Never

☐ Occasionally

☐ Sometimes

☐ Often

## 6. Consent Statement

By ticking the below check boxes I confirm that,

*Check all that apply.*

☐ I have read and understood the purpose and details of this study.

☐ I consent to perform the specified gestures under supervision.

☐ I understand that my skeletal point sequence data will be stored as numpy arrays and no videos of me will be stored.

☐ I understand that my data will only be used for research purposes and will not be shared with any external parties.

☐ I understand that I can withdraw my participation at any time without penalty.

## Evaluation of LSTM

Please fill the following with your honest opinion

7. Did the system recognize your expressions when you laughed heavily (High Laugh)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be

8. Did the system recognize your expressions when you laughed slightly (Subtle Laugh)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be

9. Did the system recognize your expressions when you were confused (Confused)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be

10. Did the system recognize your expressions when you were neutral (Neutral)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be



## 11. Comments

---

---

---

---

---

## Evaluation DTW

Please fill the following with your honest opinion

12. Did the system recognize your expressions when you laughed heavily (High Laugh)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be

13. Did the system recognize your expressions when you laughed slightly (Subtle Laugh)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be

14. Did the system recognize your expressions when you were confused (Confused)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be

15. Did the system recognize your expressions when you were neutral (Neutral)? \*

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ May be

16. Comments

---

---

---

---

---

---

This content is neither created nor endorsed by Google.

Google Forms

# Data Collection Form for VirExp: Automated Emotion and Gestures for Virtual Collaborative Environment

**Research Conducted by:**

P. G. M. N. Pupulewatte, W. U. C. M. Perera, S. M. R. L. A. Senadheera

University of Colombo School of Computing

**Title of Research Project:**

VirExp: Automated Emotion and Gestures for Virtual Collaborative Environment

*\* Indicates required question*

## General Information

1. Participant ID

---

2. Date

---

*Example: January 7, 2019*

3. Age (in years) :

*Mark only one oval.*

☐ < 18

☐ 18–25

☐ 26–35

☐ 36–50

☐ > 50

## 4. Gender:

*Mark only one oval.*

- ☐ Male
- ☐ Female
- ☐ Prefer not to say
- ☐ Other: \_\_\_\_\_

## 5. Familiarity with Virtual Collaborations in Virtual Spaces:

*Mark only one oval.*

- ☐ High
- ☐ Medium
- ☐ Low

### Data Collection Details

## 6. Have you participated in similar research before?

*Mark only one oval.*

- ☐ Yes
- ☐ No
- ☐ Maybe

## Explanation of the Expressions captured

### High Laugh

- Visible facial changes: Wide-open mouth, raised cheeks, crinkled eyes, and visible teeth.
- Body movements: Forward tilting, possible head bobbing, and exaggerated gestures (e.g., clapping, extended arms or touching the stomach).
- Eg - <https://images.app.goo.gl/b3DHAp8ky8WFzTsH9> , <https://images.app.goo.gl/i8xgoWdcccU4ktz8V9> , <https://images.app.goo.gl/cGyXLmc7x754ppqx6> , [https://www.youtube.com/watch?v=1rVuRhpLScI&ab\\_channel=9MillionMemes](https://www.youtube.com/watch?v=1rVuRhpLScI&ab_channel=9MillionMemes)

### Subtle Laugh

- Facial expressions: Slight smile, mildly raised cheeks, minimal eye crinkling.
- Body movements: Slight head nodding or tilting, minimal upper body movement.
- Eg <https://images.app.goo.gl/cEP8mEN5h4CpE37e8>

### Surprised

- Facial expressions: Raised eyebrows, wide-open eyes, slightly open mouth.
- Body movements: Quick backward movement, sudden head or hand gestures.
- Eg- <https://images.app.goo.gl/k3ra6t5ajSMg1t6w6> , <https://images.app.goo.gl/3V255pE2F3Q97EN47> , <https://images.app.goo.gl/f49gdtuzVtWXdBa48> , <https://images.app.goo.gl/MsXA9yBafpXHxFKu9> ,

### Neutral

- Facial expressions: Relaxed features, no strong muscle activation in the face.
- Body movements: Minimal movement, relaxed posture.
- Eg- <https://images.app.goo.gl/Mj6EjohsXxetXsqo6>

## 7. Additional Comments

---

---

---

---

---

## Ethical Considerations & Submission

By submitting this form, you acknowledge that your data will be collected and used for research purposes.

8. Do you agree to anonymized storage and withdrawal rights? \*

*Mark only one oval.*

☐ Yes

☐ No

---

This content is neither created nor endorsed by Google.

Google Forms

# Evaluation Form for VirExp: Automated Emotion and Gestures for Virtual Collaborative Environment - LSTM

**Research Conducted by:**

P. G. M. N. Pupulewatte, W. U. C. M. Perera, S. M. R. L. A. Senadheera

University of Colombo School of Computing

**Title of Research Project:**

VirExp: Automated Emotion and Gestures for Virtual Collaborative Environment

*\* Indicates required question*

## Ethical Considerations & Submission

By participating in this study, you agree to:

1. The collection and analysis of your expression data for academic research purposes
2. Anonymous storage of all personal information (your identity will not be linked to your responses)
3. The right to withdraw your participation at any time without penalty

1. *Do you consent to these terms? \**

*Mark only one oval.*

- ☐ Yes, I fully agree to participate under these conditions
- ☐ No, I do not wish to participate

## General Information

2. Participant ID

---

## 3. Email

---

## 4. Date

---

*Example: January 7, 2019*

## 5. Age (in years) :

*Mark only one oval.*

☐ < 18

☐ 18–25

☐ 26–35

☐ 36–50

☐ > 50

## 6. Gender:

*Mark only one oval.*

☐ Male

☐ Female

☐ Prefer not to say

☐ Other: \_\_\_\_\_

## 7. Have you participated in similar research before? \*

*Mark only one oval.*

☐ Yes

☐ No

☐ Maybe



8. Familiarity with Virtual Collaborations in Virtual Spaces. This include virtual meetings (via Zoom, Google Meet, Microsoft Teams), Metaverse environments such as Decentraland, Sandbox, RecRoom and PartySpace. \*

*Mark only one oval.*

- ☐ High
- ☐ Medium
- ☐ Low

## Evaluation

This study focuses specifically on evaluating four archetypal expressions (High Laugh, Subtle Laugh, Surprise, and Neutral) rather than assessing a comprehensive range of emotional states.

In this study, you will participate in four structured trials - one for each of our target expressions (High Laugh, Subtle Laugh, Surprise, and Neutral).

Each trial will follow the same format:

- You'll be asked to perform one specific expression
- Our system will detect and translate it to avatar animation
- You'll then evaluate how accurately both the detection and representation matched your intention

## Trial 1 : High Laugh

### High Laugh

- Visible facial changes: Wide-open mouth, raised cheeks, crinkled eyes, and visible teeth.
- Body movements: Forward tilting, possible head bobbing, and exaggerated gestures (e.g., clapping, extended arms or touching the stomach).

### High Laugh - Example 1



### High Laugh - Example 2



### High Laugh - Example 3



9. Which expression were you trying to convey?

Mark only one oval.

- ☐ High Laugh
- ☐ Subtle laugh
- ☐ Surprise
- ☐ Neutral

10. How accurately did the system detect your expression? \*

Mark only one oval.

1 2 3 4 5

Not ☐ ☐ ☐ ☐ ☐ Perfectly

11. If rating  $\leq 3$ :

Briefly describe the discrepancy you observed:

---

---

---

---

---

12. How confident are you in your rating?

Mark only one oval.

1 2 3 4 5

Low ☐ ☐ ☐ ☐ ☐ High

13. *How well did the avatar's animation match your intended expression ? \**

*Mark only one oval.*

1   2   3   4   5

No ☐ ☐ ☐ ☐ ☐ Perfectly

14. *If rating  $\leq 3$ :*

*What did the avatar display instead?*

---

---

---

---

---

15. *How confident are you in your rating?*

*Mark only one oval.*

1   2   3   4   5

Low ☐ ☐ ☐ ☐ ☐ High

## Trial 2 : Subtle Laugh

### Subtle Laugh

- Facial expressions: Slight smile, mildly raised cheeks, minimal eye crinkling.
- Body movements: Slight head nodding or tilting, minimal upper body movement.

## Subtle Laugh Example



16. Which expression were you trying to convey?

Mark only one oval.

- ☐ High Laugh
- ☐ Subtle laugh
- ☐ Surprise
- ☐ Neutral

17. How accurately did the system detect your expression? \*

Mark only one oval.

1   2   3   4   5

Not ☐ ☐ ☐ ☐ ☐ Perfectly

18. If rating  $\leq 3$ :

Briefly describe the discrepancy you observed:

---

---

---

---

---

19. *How confident are you in your rating?*

*Mark only one oval.*

	1	2	3	4	5	
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	High

20. *How well did the avatar's animation match your intended expression ? \**

*Mark only one oval.*

	1	2	3	4	5	
No ε	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Perfectly

21. *If rating  $\leq 3$ :*

*What did the avatar display instead?*

---

---

---

---

---

22. *How confident are you in your rating?*

*Mark only one oval.*

	1	2	3	4	5	
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	High

## Trial 3 : Surprised

### Surprised

- Facial expressions: Raised eyebrows, wide-open eyes, slightly open mouth.
- Body movements: Quick backward movement, sudden head or hand gestures.

### Surprised - Example 1



### Surprised - Example 2



**Surprised - Example 3****Surprised - Example 4**

23. Which expression were you trying to convey?

Mark only one oval.

- ☐ High Laugh
- ☐ Subtle laugh
- ☐ Surprise
- ☐ Neutral



24. *How accurately did the system detect your expression? \**

*Mark only one oval.*

1   2   3   4   5

Not ☐ ☐ ☐ ☐ ☐ Perfectly

25. *If rating  $\leq 3$ :*

*Briefly describe the discrepancy you observed:*

---

---

---

---

---

26. *How confident are you in your rating?*

*Mark only one oval.*

1   2   3   4   5

Low ☐ ☐ ☐ ☐ ☐ High

27. *How well did the avatar's animation match your intended expression ? \**

*Mark only one oval.*

1   2   3   4   5

No ☐ ☐ ☐ ☐ ☐ Perfectly

28. *If rating  $\leq 3$ :*

*What did the avatar display instead?*

---

---

---

---

---

29. *How confident are you in your rating?*

*Mark only one oval.*

	1	2	3	4	5	
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	High

#### Trial 4 : Neutral

##### Neutral

- Facial expressions: Relaxed features, no strong muscle activation in the face.
- Body movements: Minimal movement, relaxed posture.

##### Neutral Example



30. *Which expression were you trying to convey?*

*Mark only one oval.*

- ☐ High Laugh
- ☐ Subtle laugh
- ☐ Surprise
- ☐ Neutral

31. *How accurately did the system detect your expression? \**

*Mark only one oval.*

1   2   3   4   5

Not ☐ ☐ ☐ ☐ ☐ Perfectly

32. *If rating  $\leq 3$ :*

*Briefly describe the discrepancy you observed:*

---

---

---

---

---

33. *How confident are you in your rating?*

*Mark only one oval.*

1   2   3   4   5

Low ☐ ☐ ☐ ☐ ☐ High

34. *How well did the avatar's animation match your intended expression ? \**

*Mark only one oval.*

	1	2	3	4	5	
No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Perfectly

35. *If rating  $\leq 3$ :*  
*What did the avatar display instead?*

---



---



---



---



---

36. *How confident are you in your rating?*

*Mark only one oval.*

	1	2	3	4	5	
Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	High

### Real-time Performance

37. How responsive was the system in recognizing and displaying your expressions? \*

*Mark only one oval.*

	1	2	3	4	5	
Sign	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Instantaneous

38. Did the system maintain consistent performance during prolonged use? \*

*Mark only one oval.*

1 2 3 4 5

Unu ☐ ☐ ☐ ☐ ☐ Yes, performance remained consistent

39. Were there any interruptions or delays in the avatar's expression feedback? \*

*Mark only one oval.*

1 2 3 4 5

Majr ☐ ☐ ☐ ☐ ☐ None

### Collaborative Interaction

40. Do you think that the system would enhance your collaborative experience? \*

*Mark only one oval.*

1 2 3 4 5

Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

41. How intuitive was the system for facilitating communication? \*

*Mark only one oval.*

1 2 3 4 5

Not ☐ ☐ ☐ ☐ ☐ Very Intuitive

42. Would you prefer using this kind of a system in future virtual collaborative sessions? \*

*Mark only one oval.*

- ☐ Yes
- ☐ Maybe
- ☐ No



---

This content is neither created nor endorsed by Google.

Google Forms

## LSTM Code Snippet

```
model = Sequential()
model.add(LSTM(64,
              return_sequences=True,
              activation='tanh',
              dropout=0.2,
              recurrent_dropout=0.2,
              input_shape=(30, 1662)))
model.add(LSTM(128,
              return_sequences=True,
              activation='tanh',
              dropout=0.2,
              recurrent_dropout=0.2))
model.add(LSTM(64,
              return_sequences=False,
              activation='tanh',
              dropout=0.2,
              recurrent_dropout=0.2))
model.add(Dropout(0.3))
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.3))
model.add(Dense(32, activation='relu'))
model.add(Dense(actions.shape[0], activation='softmax'))
```

## DTW Code Snippet

```
def dtw_distance(seq1, seq2):
    distances = []
    for i in range(len(seq1[0, :])):
        dist = dtw.distance_fast(seq1[:, i], seq2[:, i])
        distances.append(dist)
    return np.mean(distances)
```

## Transformer Code Snippets

```
class TransformerEncoderLayer(tf.keras.layers.Layer):
    def __init__(self, embed_dim, num_heads, hidden_dim, dropout=0.1, **kwargs):
        super().__init__(**kwargs)
        self.embed_dim = embed_dim
        self.num_heads = num_heads
        self.hidden_dim = hidden_dim
        self.dropout = dropout

        self.attention = MultiHeadAttention(
            num_heads=num_heads, key_dim=embed_dim // num_heads
        )
        self.dense1 = Dense(hidden_dim, activation="relu")
        self.dense2 = Dense(embed_dim)
        self.layernorm1 = LayerNormalization(epsilon=1e-6)
        self.layernorm2 = LayerNormalization(epsilon=1e-6)
        self.dropout1 = Dropout(dropout)
        self.dropout2 = Dropout(dropout)

    def call(self, inputs, training):
        attention_output = self.attention(inputs, inputs)
        attention_output = self.dropout1(attention_output, training=training)
        out1 = self.layernorm1(inputs + attention_output)
        ffn_output = self.dense1(out1)
        ffn_output = self.dense2(ffn_output)
        ffn_output = self.dropout2(ffn_output, training=training)
        return self.layernorm2(out1 + ffn_output)

    def get_config(self):
        config = super().get_config()
        config.update({
            "embed_dim": self.embed_dim,
            "num_heads": self.num_heads,
            "hidden_dim": self.hidden_dim,
            "dropout": self.dropout,
        })
        return config
```

```

class GestureTransformer(tf.keras.Model):
    def __init__(
        self,
        num_classes,
        num_heads,
        num_layers,
        hidden_dim,
        embed_dim,
        dropout,
        **kwargs # <- Important for TF to pass trainable/dtype
    ):
        super().__init__(**kwargs) # Pass kwargs to parent
        self.num_classes = num_classes
        self.num_heads = num_heads
        self.num_layers = num_layers
        self.hidden_dim = hidden_dim
        self.embed_dim = embed_dim
        self.dropout = dropout

        self.projection = Dense(embed_dim)
        self.encoder_layers = [
            TransformerEncoderLayer(embed_dim, num_heads, hidden_dim, dropout)
            for _ in range(num_layers)
        ]
        self.fc = Dense(num_classes, activation="softmax")

    def call(self, x, training=False):
        x = self.projection(x)
        for encoder_layer in self.encoder_layers:
            x = encoder_layer(x, training=training)
        x = tf.reduce_mean(x, axis=-1)
        x = self.fc(x)
        return x

    def get_config(self):
        config = super().get_config()
        config.update({
            "num_classes": self.num_classes,
            "num_heads": self.num_heads,
            "num_layers": self.num_layers,
            "hidden_dim": self.hidden_dim,
            "embed_dim": self.embed_dim,
            "dropout": self.dropout,
        })
        return config

```