Development Process of VR-Based Cognitive Assessments for Children with Developmental Delays

Chomyong Kim[1], Eun Young Kim[2], Yunyoung Nam[1,3*]

[1]*Emotional and Intelligence Child Care System Convergence Research Center, Soonchunhyang University, Asan 31538, Republic of Korea*

[2]*Department of Occupational Therapy, Soonchunhyang University, Asan 31538, Republic of Korea*

[3]*Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Republic of Korea*

*Contact: First.Author@sfu.ca, phone +1-778 782 0000

Abstract— **Virtual reality (VR) has been utilized in cognitive assessments, yet research targeting young children remains limited. This study addresses this gap by developing a VR-based cognitive assessment for three-year-olds using the Meta Quest Pro with eye gaze tracking. The process began with child development experts creating a detailed script and video materials for tasks such as object recognition, colour identification, counting, and comparisons. Collaborating with animation professionals, the team designed familiar avatars and interactive objects in Unity, ensuring a realistic daycare-like VR environment. The tester avatar's animations were achieved through motion capture. This interdisciplinary effort highlights the potential of VR in early childhood cognitive assessment and paves the way for further research in this field.**

I. Introduction

Virtual reality (VR) technology has rapidly evolved, providing immersive environments that replicate real-world experiences. Widely applied in healthcare, education, and training, VR effectively creates controlled spaces that enhance engagement and behavior monitoring [1]. In healthcare, VR has been used to treat psychological conditions such as phobias, PTSD, and anxiety, offering safe and repeatable therapeutic environments [2]. VR is also valuable in educational training for high-risk fields like surgery and aviation, improving skill development and reducing errors [3].

In cognitive assessments, VR's multi-sensory, immersive experiences enable researchers to observe emotional, cognitive, and behavioral responses under controlled conditions, enhancing assessment accuracy [4], [5], [6]. VR's potential in pediatric healthcare is notable, particularly for developmental disorder assessments. Studies show VR training programs can improve executive functions in children with ADHD [7]. VR classroom simulations have also enhanced attention disorder evaluations by recreating real-life classroom stimuli [8]. Furthermore, VR tools have been effective in improving social interaction in children with autism spectrum disorder [9].

Despite VR's promise, most pediatric assessment studies have targeted children aged five and above [10]. Since early diagnosis and intervention are crucial to prevent developmental delays from progressing into severe intellectual disabilities [11], there is an urgent need for tools designed for younger children. Early childhood (ages 0–5) is a vital period for cognitive development, making targeted assessments essential.

This study addresses this gap by developing a VR-based cognitive assessment system specifically for three-year-old children [12]. At this age, children can recognize avatars as self-representations and interact meaningfully with virtual environments [13]. The system assesses cognitive functions such as object recognition, categorization, color matching, memory, and numerical understanding — key markers for developmental progress [14].

The assessment leverages the Meta Quest 3 with hand-tracking technology, allowing children to interact naturally with objects without controllers. This method improves accessibility for younger children who may struggle with conventional VR controllers. Designed to resemble daycare settings, the VR environment enhances ecological validity by mirroring real-life experiences.

Recent upgrades to the Meta Quest Pro with integrated eye-tracking technology enable a deeper evaluation of cognitive abilities. Gaze tracking assesses mental function through looking-time paradigms, which measure visual responses when paired with auditory stimuli. Studies show that gaze patterns can reveal cognitive understanding in infants and predict future IQ levels [15], [16]. Eye-tracking is particularly valuable for children with limited verbal skills, as gaze patterns often provide clearer insights than verbal behavior [17].

Language skills are closely linked to cognition, making verbal responses a vital assessment metric. For example, naming tasks can evaluate expressive vocabulary, a strong indicator of cognitive ability [18], [19], [20]. Additionally, analyzing word retrieval speed can help identify intellectual impairments, as children with developmental delays often exhibit slower object-naming speeds [22], [23].

Motion tracking further enhances assessment by capturing upper-body movement data as children reach, grasp, or place objects. This data is critical for identifying motor development issues, which are often linked to intellectual disabilities [24], [25]. Studies indicate that reduced manual dexterity correlates with the severity of intellectual impairment [26].

The development of this VR-based assessment tool involved multidisciplinary collaboration across occupational therapy, developmental psychology, computer science, and digital animation. This collaboration ensured the tool's developmental appropriateness and scientific rigor, reinforcing its potential as an effective early childhood assessment solution.

This innovative VR cognitive assessment tool for three-year-olds offers a novel approach to early diagnosis and intervention. By combining advanced VR technology with insights from child development experts, this tool holds promise for improving developmental outcomes for children with and without disabilities. Further research is encouraged to expand its applications and validate its effectiveness across diverse populations.

II. Research Procedure

The development of the VR cognitive assessment tool followed a structured process (Figure 1). It began with script composition, where child development experts designed scenarios featuring tasks such as memory recall, object recognition, and classification to ensure developmental appropriateness and engagement.

Next, non-player characters (NPCs), including a teacher and child avatars, were designed to appear familiar and approachable, enhancing the comfort and immersion of young participants.

Object modeling followed, where interactive toys were created and programmed with physics engines to simulate realistic responses, maintaining ecological validity.

The virtual environment was designed to resemble daycare and play settings, with customizable elements like lighting and object arrangement. An administration UI was developed to enable researchers to control the environment and collect data based on child interactions.

Finally, animation was achieved using advanced motion capture technology to animate NPCs with natural and lifelike movements, enhancing the overall immersive experience for children.

III. Research Design

A. Script Composition

A comprehensive script was designed to create cognitive assessment tasks targeting key developmental areas: object concepts, color concepts, and numerical understanding. The tasks included activities like remembering objects, naming items, matching colors, and counting.

For object concepts, tasks assessed memory, recognition, and classification skills. Children identified hidden objects, named familiar items, matched similar objects, and grouped items into categories like fruits, animals, or vehicles.

Color concept tasks involved naming, matching, and classifying colors. Numerical concept tasks included counting objects, understanding the concept of “one,” and making quantity comparisons.

The script detailed avatar positioning, object placement, and data collection points to guide developers in creating accurate VR content. Video materials were produced to ensure consistent animations and interactions. The script underwent multiple revisions with input from psychologists and VR developers to ensure developmental appropriateness and engagement for three-year-olds.

B. Development tools and Technologies

The development of the VR cognitive assessment tool utilized advanced technologies to enhance usability and engagement for young children. The Meta Quest Pro's HandTracking feature enabled natural hand gestures, eliminating the need for controllers and making interactions easier for children.

A physics engine was integrated to create realistic object responses, ensuring toys and balls moved naturally. To prevent distractions from excessive movement, an interaction range was defined, automatically returning objects to their original position if they moved too far.

To support children with developing motor skills, collider functionality was added, with dynamically adjusted sizes to facilitate smoother interactions. Additionally, a dissolve effect was applied for object appearance and disappearance, ensuring smooth transitions to minimize distraction and maintain focus.

C. Content Development

The content development process involved creating a child-friendly VR environment with carefully designed elements to ensure engagement and realism.

The VR environment featured four NPCs: one teacher (tester) and three children (two girls and one boy). To enhance familiarity and comfort, the teacher NPC resembled a typical Korean nursery teacher, while child NPCs were designed with simple, warm visuals inspired by Korean children’s animations. Features like slightly enlarged heads, softer jawlines, and smaller eyes improved relatability and ensured a child-friendly appearance.

For object modeling, interactive items such as toy animals, vehicles, and balls were designed with realistic textures. Colliders were integrated to enhance responsiveness, ensuring smooth interactions through physics-based reactions.

The VR environment simulated a daycare setting with neutral colors and familiar layouts to minimize distractions. A modular design allowed adjustments in lighting and object placement. The Administrator UI enabled content control, data collection, and customization of the child’s visual perspective.

Animations were developed using motion capture, performed by a Ph.D. occupational therapist for accuracy. Each movement was refined through skeletal structuring, motion adjustments, and synchronized with voice recordings to ensure natural interactions.

D. Data Collection and Analysis Module

The The Meta Quest Pro VR headset was utilized for both content playback and data collection, capturing data such as gaze tracking, facial expressions, head movement, and hand gestures to assess developmental delays.

Gaze Pattern Analysis involved comparing the visual attention patterns of typically developing children and those with developmental delays.

The Eye Gaze Observer Module collected and normalized gaze data in Unity's environment, ensuring consistent results across screen resolutions. It also enabled object-specific gaze tracking using bounding box coordinates.

The Gaze Tracking Module identified Areas of Interest (AOIs) to measure children's focus and engagement.

The Blink Correction Module corrected abrupt gaze shifts caused by blinking by averaging surrounding gaze points, ensuring smoother data flow.

The Gaze Analysis Module measured various cognitive metrics, including:

- **Time to First Fixation** (time taken to focus on a target),
- **Dwell Time** (total time spent fixating on an object),
- **Fixation Sequences** (detailed gaze movement patterns),
- **Revisit Counts** (frequency of returning gaze to specific objects),
- **First Fixation Duration** and **Average Fixation Duration**, providing insights into attention span and cognitive focus.

These data points, along with heatmap visualizations, provided valuable insights for identifying developmental patterns and distinguishing between typically developing children and those with delays.

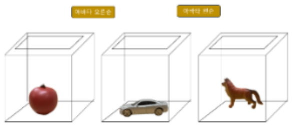
IV. RESULT

A. Script

An initial draft script was created to outline the core concepts and tasks for the cognitive assessment, including activities like object classification and memory recall. This script, developed by a child assessment expert, was refined in collaboration with program developers to ensure it was both age-appropriate and engaging. Table 1 provides examples from the "Classifying Objects" task.

.

TABLE I
SCRIPT EXAMPLES OF SCENE CONFIGURATIONS FOR "CLASSIFYING OBJECTS" COGNITIVE ASSESSMENT TASKS

Item	[Task example] Classifying Objects
Scene description	#12 In the same background as #11, all four boxes disappear, and three transparent boxes each containing one fruit, one vehicle, or one animal appear on the table. - The tester avatar places both palms slightly above the top of the transparent boxes and looks at the child, saying the line (1). - the avatar lowers her arms and says (2)
Tester avatar's instruction line (and action)	(1) "There are boxes with toys here." (2) "I will put toys in the box first."
Example picture	
Responses to be collected from children	- Gaze [fixation time on each box]

Scene transition time	2 s

The video material recorded according to the script were useful because they could show the final VR situation as a visual image. Figure 4 shows scene #13 in the video materials.

B. Content Development

1) NON-PLAYER CHARACTERS

The NPCs were designed to be familiar and approachable for young children, featuring one adult tester and three children (one boy and two girls). This intentional design choice aimed to foster engagement and provide a sense of comfort for the participants. To achieve realistic animations, various facial expressions and movements were created using Blendshapes, while lip-sync functionality was implemented to synchronize the NPCs' speech with their mouth movements. These design elements contributed to a dynamic and immersive environment, enhancing the overall experience for children. Figure 1 illustrates the developed NPCs, emphasizing their simplicity and detailed animations.

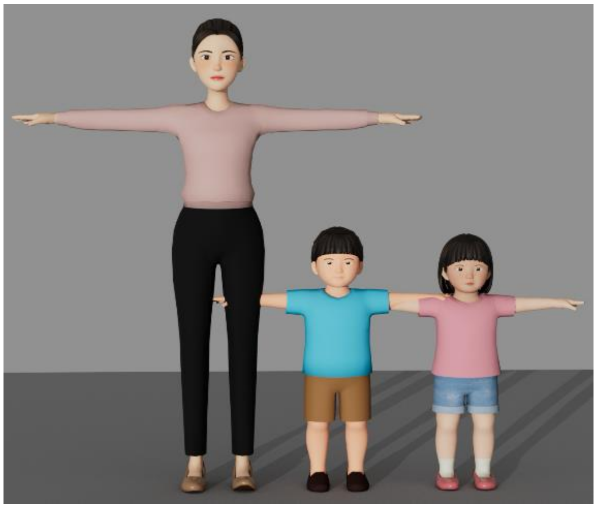


Fig. 1 Developed tester NPC and boy and girl NPC

2) OBJECT MODELING

All objects utilized in the study were derived from the testing aids employed during the script development phase and supplemented with open-source resources such as the Unity Asset Store and Sketchfab (Figure 6). The selection of these objects was directly overseen by a professional responsible for designing the assessment tasks. Particular attention was given to the familiarity and complexity of the objects, as these factors were crucial in evaluating children. Some objects obtained from external sources were originally designed for games, often featuring unrealistic visual effects such as intense lighting or flashing effects. In such instances, these components were removed to maintain consistency. For ultra-realistic high-poly objects, unnecessary layers were simplified to reduce the polygon count, minimizing both game size and potential distractions. Furthermore, to avoid startling children by having objects abruptly disappear, a dissolve shader effect was incorporated — a familiar visual cue often seen in popular media.

3) VIRTUAL ENVIRONMENT

In the study's initial phase, the virtual environment was designed to resemble familiar settings for South Korean children aged 3 to 6, such as daycare centers, nurseries, or homes. The environment featured light beige and soft gray tones with bright lighting to ensure a well-lit, neutral atmosphere. To minimize distractions, the background and wallpaper were kept simple, and windows were designed to block outside views. A light-colored table displayed the objects, which were either purchased or developed based on toys from the pilot study. A direct overhead light illuminated the objects to help maintain focus. Unity's default URP shaders and simple textures were used to reduce visual distractions.The child's gaze range and points of interest (Tracking-Space) were predefined, with gaze deviation recorded as a negative value. Audio cues were linked to the tester NPC, with spatial sound effects implemented to reflect the child's head movements for enhanced realism.

4) ANIMATING

Following comprehensive preparatory tasks, motion capture for cognitive content production was conducted. Each motion was recorded 3 to 5 times, with the session lasting approximately 3 to 4 hours. During content development, adjustments were necessary for several motions. Since additional filming was not feasible, animations were refined by modifying the position, rotation, and transform values of each joint, resulting in over six iterations of animation production.

A total of 32 animation clips were created, demonstrating tasks such as picking up a ball, showing it, or placing it in a box. Each clip was approximately 15 seconds long. The finalized animations were integrated into Unity and synchronized with voice recordings from a professional actor, ensuring a cohesive user experience.

C. *Data Collection and Analysis Module*

The gaze tracking module effectively captured and corrected missing data caused by blinking, achieving 92.7% accuracy in blink detection. This improved the precision of gaze movement analysis. To ensure data reliability, blink occurrences were estimated, and missing gaze data were corrected using interpolation. The system achieved a 7.3% error in blink estimation, indicating high accuracy.

Gaze coordinates during blinks were ignored and replaced with the average of preceding and following points, reducing abrupt changes and improving gaze movement reconstruction (Figure 2).

Additionally, object size was tested for its impact on gaze analysis. Results showed that smaller objects reduced gaze point accuracy, highlighting the importance of appropriate object sizing in VR content for reliable tracking. Analysis of fixation and saccade patterns further revealed that children with developmental delays showed shorter fixations and more erratic gaze patterns, suggesting challenges in sustained attention and visual processing [7].

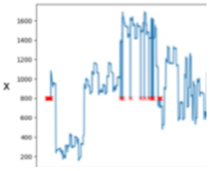
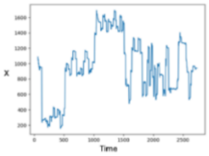
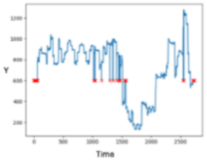
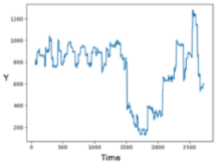
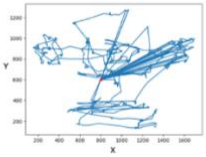
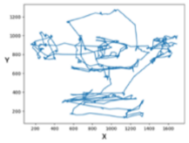
	Before Correction	After Correction
Visualization of Gaze Coordinates on the X-Axis Over Time		
Visualization of Gaze Coordinates on the Y-Axis Over Time		
Visualization of X and Y Coordinates		

Fig. 2 Visualization of gaze movement: before correction (left) and after correction (right)

V. Discussion

The VR cognitive development assessment (Figure 10) was developed through a multidisciplinary process involving script composition, motion capture, avatar creation, Unity configuration, motion application, and content refinement. This collaboration integrated expertise from child occupational therapy, developmental psychology, digital animation, software engineering, and computer science to ensure the content was both developmentally appropriate and technically advanced.

During script composition, detailed information and multiple revisions ensured content validity. Drawing from previous computer-based language assessment templates helped structure scenarios to align with developmental goals. Visual aids such as video materials and character models were instrumental in improving interdisciplinary communication and streamlining development. Using an actor as a reference for avatar design enhanced the relatability and friendliness of NPCs for child participants.

Research indicates that computerized tasks effectively assess young children's cognitive functions, with tablet-based assessments successfully measuring memory, attention, and categorization skills. While VR offers immersive and interactive environments with greater ecological validity, no prior studies have explored VR-based assessments specifically for three-year-olds. This study addresses that gap, representing a significant advancement in early childhood cognitive assessment.

The developed VR assessment tool has key implications. It enables early detection of developmental delays, allowing timely interventions that can improve long-term outcomes. Its immersive nature enhances assessment accuracy compared to traditional methods. Furthermore, this study extends VR’s applications into pediatric healthcare and education, demonstrating its potential for broader social impact. The project also highlights the value of interdisciplinary collaboration in creating innovative assessment solutions.

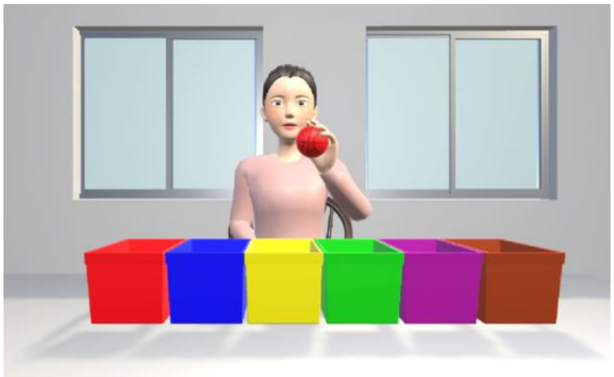


Fig. 3 Scene from the final assessment simulation with all developed NPCs, objects and animations

Despite its success, the study has limitations. It lacks extensive validation comparing VR outcomes with traditional methods, posing challenges in establishing reliability. Technological constraints, such as hand-tracking accuracy and VR equipment costs, may limit accessibility. Future research should focus on refining interaction methods, exploring cost-effective hardware solutions, and expanding assessment modules to evaluate language, social interaction, and motor skills. Integrating AI-driven personalization could further enhance adaptive assessments tailored to individual children's needs.

In conclusion, while this study marks a major step in VR-based cognitive assessment development, addressing these limitations through validation, expanded functionality, and improved accessibility will be crucial for broader implementation and impact.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund.

References

[1] P. M. G. Emmelkamp and K. Meyerbröker, "Virtual reality therapy in mental health," *Annual Review of Clinical Psychology*, vol. 17, no. 1, pp. 495–519, Mar. 2021. [Online]. Available: https://doi.org/10.1146/annurev-clinpsy-081219-115923

[2] S. K. Renganayagalu, S. C. Mallam, and S. Nazir, "Effectiveness of VR head-mounted displays in professional training: A systematic review," *Technology, Knowledge and Learning*, vol. 26, no. 1, pp. 1–43, Mar. 2021. [Online]. Available: https://doi.org/10.1007/s10758-020-09489-9

[3] B. Wu, X. Yu, and X. Gu, "Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis," *British Journal of Educational Technology*, vol. 51, no. 6, pp. 1991–2005, Nov. 2020. [Online]. Available: https://doi.org/10.1111/bjet.13023

[4] P. Kourtesis, S. Collina, L. A. Doumas, and S. E. MacPherson, "Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An immersive virtual reality neuropsychological battery with enhanced ecological validity," *Journal of the International Neuropsychological Society*, vol. 27, no. 2, pp. 181–196, Feb. 2021. [Online]. Available: https://doi.org/ 10.1017/S1355617720000764

[5] E. Bozkir, D. Geisler, and E. Kasneci, "Assessment of driver attention during a safety-critical situation in VR to generate VR-based training," in *Proceedings of the ACM Symposium on Applied Perception (SAP)*, Barcelona, Spain, Sept. 2019, pp. 1–5. [Online]. Available: https://doi.org/10.1145/3343036.3343138

[6] T. Segawa et al., "Virtual reality (VR) in assessment and treatment of addictive disorders: A systematic review," *Frontiers in Neuroscience*, vol. 13, p. 1409, Dec. 2020. [Online]. Available: https://doi.org/ 10.3389/fnins.2019.01409

[7] A. Bashiri, M. Ghazisaeedi, and L. Shahmoradi, "The opportunities of virtual reality in the rehabilitation of children with attention deficit hyperactivity disorder: A literature review," *Korean Journal of Pediatrics*, vol. 60, no. 11, pp. 337–343, Nov. 2017. [Online]. Available: https://doi.org/10.3345/kjp.2017.60.11.337

[8] Y. Iriarte, U. Diaz-Orueta, E. Cueto, P. Irazustabarrena, F. Banterla, and G. Climent, "AULA—Advanced virtual reality tool for the assessment of attention: Normative study in Spain," *Journal of Attention Disorders*, vol. 20, no. 6, pp. 542–568, Sept. 2016. [Online]. Available: https://doi.org/10.1177/1087054712465335

[9] R. Bradley and N. Newbutt, "Autism and virtual reality head-mounted displays: A state of the art systematic review," *Journal of Enabling Technologies*, vol. 12, no. 3, pp. 101–113, Sept. 2018. [Online]. Available: https://doi.org/10.1108/JET-01-2018-0004

[10] E. Seesjärvi, J. Puhakka, E. T. Aronen, A. Hering, S. Zuber, L. Merzon, M. Kliegel, M. Laine, and J. Salmi, "EPELI: A novel virtual reality task for the assessment of goal-directed behavior in real-life contexts," *Psychological Research*, vol. 87, no. 6, pp. 1899–1916, Sept. 2023. [Online]. Available: https://doi.org/10.1007/s00426-022-01770-z.

[11] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., Arlington, VA, USA: American Psychiatric Publishing, 2013. [Online]. Available: https://doi.org/10.1176/appi.books.9780890425596

[12] M. Hong, E. Y. Kim, Y. Nam, C. Kim, and J. Kim, "Development of metaverse assessment items for early screening of developmental disabilities," *The Journal of Korean Academy of Sensory Integration*, vol. 22, no. 2, pp. 41–54, Apr. 2024. [Online]. Available: https://doi.org/ 10.18064/JKASI.2024.22.2.41

[13] T. Suddendorf and D. L. Butler, "The nature of visual self-recognition," *Trends in Cognitive Sciences*, vol. 17, no. 3, pp. 121–127, Mar. 2013. [Online]. Available: https://doi.org/10.1016/j.tics.2013.01.004

[14] E. Weiner and D. Sanchez, "Cognitive ability in virtual reality: Validity evidence for VR game-based assessments," *International Journal of Selection and Assessment*, vol. 28, no. 3, pp. 215–235, Sept. 2020. [Online]. Available: https://doi.org/10.1111/ijsa.12295

[15] R. M. Golinkoff, W. Ma, L. Song, and K. Hirsh-Pasek, "Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned?" *Perspectives on Psychological Science*, vol. 8, no. 3, pp. 316–339, May 2013. [Online]. Available: https://doi.org/10.1177/1745691613484936

[16] S. A. Rose, J. F. Feldman, J. J. Jankowski, and R. Van Rossem, "Information processing from infancy to 11 years: Continuities and prediction of IQ," *Intelligence*, vol. 40, no. 5, pp. 445–457, Sept.–Oct. 2012. [Online]. Available: https://doi.org/10.1016/j.intell.2012.05.007

[17] E. Vakil, H. Lifshitz, D. Tzuriel, I. Weiss, and Y. Arzuoan, "Analogies solving by individuals with and without intellectual disability: Different cognitive patterns as indicated by eye movements," Research in Developmental Disabilities, vol. 32, no. 2, pp. 846–856, Mar.–Apr. 2011. [Online]. Available: https://doi.org/10.1016/j.ridd.2010.08.006

[18] E. V. Clark, "How language acquisition builds on cognitive development," Trends in Cognitive Sciences, vol. 8, no. 10, pp. 472–478, Oct. 2004. [Online]. Available: https://doi.org/10.1016/j.tics.2004.08.012

[19] D. Gatt, H. Grech, and B. Dodd, "Early expressive vocabulary skills: A multi-method approach to measurement," First Language, vol. 34, no. 2, pp. 136–154, Apr. 2014. [Online]. Available: https://doi.org/10.1177/0142723714521830

[20] L. S. DeThorne, S. A. Petrill, M. E. Hayiou-Thomas, and R. Plomin, "Low expressive vocabulary: Higher heritability as a function of more severe cases," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 4, pp. 792–804, Aug. 2005. [Online]. Available: https://doi.org/10.1044/1092-4388(2005/055)

[21] M. van der Schuit, E. Segers, H. van Balkom, and L. Verhoeven, "How cognitive factors affect language development in children with intellectual disabilities," *Research in Developmental Disabilities*, vol. 32,

no. 5, pp. 1884–1894, Sept.–Oct. 2011. [Online]. Available: https://doi.org/10.1016/j.ridd.2011.03.015

[22] A. F. de Chambrier, R. S. Dessemontet, C. Martinet, and M. Fayol, "Rapid automatized naming skills of children with intellectual disability," *Heliyon*, vol. 7, no. 5, p. e06944, May 2021. [Online]. Available: https://doi.org/10.1016/j.heliyon.2021.e06944

[23] M. E. Foster, R. A. Sevcik, M. Romski, and R. D. Morris, "Effects of phonological awareness and naming speed on mathematics skills in children with mild intellectual disabilities," *Developmental Neurorehabilitation*, vol. 18, no. 5, pp. 304–316, 2015. [Online]. Available: https://doi.org/10.3109/17518423.2013.843603

[24] P. Rintala and E. M. Loovis, "Measuring motor skills in Finnish children with intellectual disabilities," *Perceptual and Motor Skills*, vol. 116, no. 1, pp. 294–303, Feb. 2013. [Online]. Available: https://doi.org/10.2466/25.10.PMS.116.1.294-303

[25] C. Maïano, O. Hue, and J. April, "Effects of motor skill interventions on fundamental movement skills in children and adolescents with intellectual disabilities: A systematic review," *Journal of Intellectual Disability Research*, vol. 63, no. 9, pp. 1163–1179, Sept. 2019. [Online]. Available: https://doi.org/10.1111/jir.12618

[26] P. J. Vuijk, E. Hartman, E. Scherder, and C. Visscher, "Motor performance of children with mild intellectual disability and borderline intellectual functioning," *Journal of Intellectual Disability Research*, vol. 54, no. 11, pp. 955–965, Nov. 2010. [Online]. Available: https://doi.org/10.1111/j.1365-2788.2010.01318.x

Automated Recognition of Parent-Child Interactions Using YOLO-Based Models

V. Sreypov[1], K. Chomyong[2], T. Sokea[1], N. Yunyoung[3*]

[1]Department of ICT Convergence, Soonchunhyang University, Asan 31538, Republic of Korea

[2]ICT Convergence Research Center, Soonchunhyang University, Asan 31538, Republic of Korea

[3]Emotional and Intelligent Child Care Convergence Center, Soonchunhyang University, Asan 31538, Republic of Korea

**Corresponding Author: Yunyoung Nam. Email: ynam@sch.ac.kr*

Abstract— **Human Interaction Recognition (HIR) is the process of detecting and interpreting interactive actions and activities among multiple participants in a specific environment. HIR has a wide range of applications in video analysis, such as security systems, sports analysis, entertainment, and the healthcare domain. The aim of this study is to propose a robust system capable of recognizing parent-child interactions using CCTV footage dataset and classifying them into a binary class: interaction or non-interaction. The dataset was collected in an indoor environment using a one-channel camera, capturing multiple actions between pairs of parents and children. The system classifies parent-child interactions based on nonverbal behaviours, including skeletal joint positions and head rotation, which together represent human activity. The YOLOv11 model was employed to generate individual skeleton data, while YOLOv8 was used for parent-child classification and head detection. DeepSORT was integrated to enable real-time object tracking. The proposed framework has been validated through extensive experiments, showing its effectiveness in detecting and classifying parent-child interactions, achieving an overall accuracy of 94.74%. The results indicate its potential for practical applications in monitoring and analysing dynamic interactions.**

I. Introduction

Parent-child interaction recognition from video surveillance is a challenging research topic in the branch of human activity recognition. Human interaction refers to how people communicate and engage with each other, including verbal and nonverbal communication through voice/motion/eye contact, facial expression, gestures, or any physical actions [20]. In this work, we present a study of nonverbal communication. It is well known that investigating parent-child interaction plays an important role in understanding child development, child behaviours, and as an indicator of the relationship between parent and child [1]. Another crucial reason is to help in investigating any delay issues of the child earlier. On the other hand, observing how children interact with their parents can provide insights into their emotional, social, and cognitive development. This can help in designing strategies to support and enhance their growth.

The effective parental involvement in children’s daily activities greatly influences their cognitive development and enhances communication. This underscores the importa nce of interventions designed to enhance parental engageme nt and promote meaningful parent-child interactions.

In a recent work, the most effective solutions are based on advanced deep learning techniques, particularly deep neural networks (DNNs) such as Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), and Long Short-Term Memory Networks (LSTMs). These techniques excel in handling various types of data, with CNNs being highly effective for image and spatial data processing, GCNs specialized for relational data and network analysis, and LSTMs particularly suited for sequential and time-series data. Human activity recognition in video surveillance can be categorized into two approaches: directly applying recognition techniques to the video dataset, or first conducting human pose estimation (i.e., skeleton detection, identified bounding boxes) on each frame of the sequence [3].

Human pose estimation is a prominent topic in current research due to its wide range of applications, such as motion capture, telepresence, and object manipulation in virtual environments. It involves identifying key joints and body parts in an image or video to accurately determine a person’s pose. The result is typically a set of keypoints or a skeleton graph, which can be used for tasks like motion analysis, action recognition, and interactive applications [4].

Likewise, the researchers in Puchała et al. [3] employed skeleton data tracking and feature extraction for human interaction classification, using either OpenPose or HRNet for pose estimation, depending on their setup. These tools provided the necessary skeleton data, which were later processed using an LSTM-based model to classify two-person interactions in video sequences.

Similarly, in [5] leveraged OpenPose for human pose estimation in their hybrid deep learning and machine learning framework for human interaction recognition in surveillance videos. The skeleton data generated by OpenPose enabled accurate recognition of human interactions.

On the other hand, [6] utilized YOLOv7 for real-time multiple object detection, combined with the FairMOT algorithm for accurate object tracking. The integration achieved notable detection accuracy, making it highly effective for tracking multiple objects in dynamic environments.

This work presents the use of the YOLOv11 model for human skeleton joints extraction, while YOLOv8 was used for real-time human classification and head detection in video sequences. These models can handle input in multiple formats, including images, videos, and real-time camera streams. We captured the interaction by segmenting from the starting point to the end of the action. This study aims to develop a robust system that enables to recognition of the parent-child interaction based on nonverbal behaviours: skeleton joints and head rotation estimation. This will allow the development of different strategies to detect more detailed interaction patterns in video surveillance in the future.

II. Proposed Methodology

A. *Overall Framework Architecture*

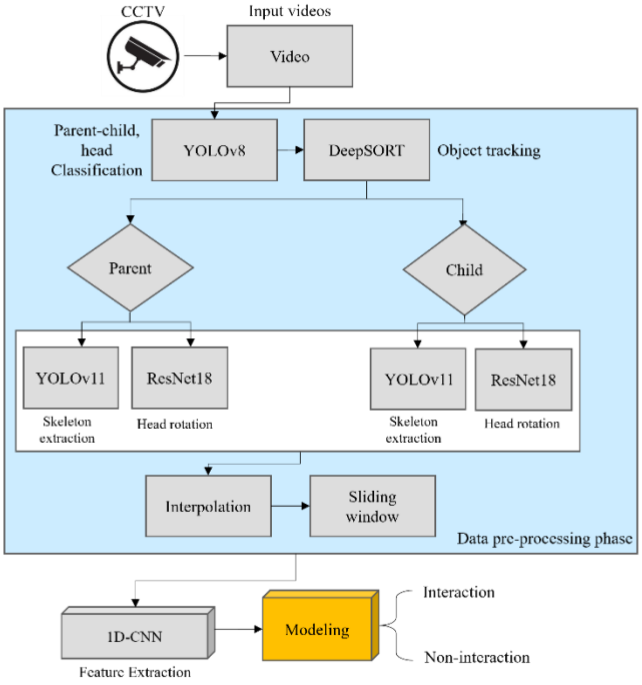


Figure 1: Architecture of interaction classification model

The proposed model in Figure 1 is designed to recognize parent-child interactions from CCTV-recorded videos using a combination of object detection, skeleton-based data analysis, and deep learning-based classification. The process involves several necessary steps, which begin with taking the sequence of video frames as input, where pre-trained YOLOv8 is employed for parent-child classification and the head detection task. DeepSORT [8, 9] was simultaneously used to continuously track each individual across frames, parents, and children.

For each detected parent and child, the model extracts two meaningful features: skeleton joint coordinates and head rotation. These features contribute to understanding interaction cues between the parent and child. In addition, these two features are later fused to enhance interaction recognition [10]. YOLOv11m-pose is used for sequence keypoints extraction, which provides 17 predefined keypoints. There are including the nose, left/right eyes, left/right ears, left/right shoulder, left/right elbow, left/right wrist, left/right hip, left/right knee, and left/right ankle for representing the body structure [11]. This provides critical information about body posture and movement. Additionally, head rotation features are extracted using a pre-trained ResNet18 [12], capturing yaw, pitch, and roll angles to represent head orientation and direction of attention. Following feature alignment through interpolation, a sliding window technique is applied during the data pre-processing phase to segment the continuous sequence of features into fixed-length overlapping windows. This is necessary for standardizing the input size for the learning stage. Each segmented window contains synchronized sequences of skeleton joint data and head rotation angles.

These windowed segments are then passed through temporal deep learning models (1D-CNN), which capture sequential patterns in the feature space. Finally, the model classifies each video segment as either “Interaction” or “Non-Interaction” based on the learned dynamics between the parent and child.

B. *Dataset*

Our video surveillances contain many activities (free play) in pairs of participants. The dataset consists of 52 parents and children, where the children in 2-6 years old engaged in a child playing environment (indoor environment). The dataset aimed at capturing natural interaction activities between a parent and their child, with their interest and enjoyment.

One-channel cameras were placed to capture interactions from a top-view angle for video data collection. The collected RGB videos range from 10 to 15 minutes in length, with a resolution of 3840 x 2160 pixels, 15 frames per second. The detailed data collection procedure is shown in Fig. 2.

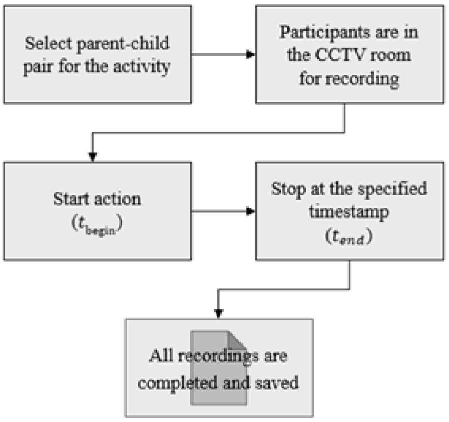
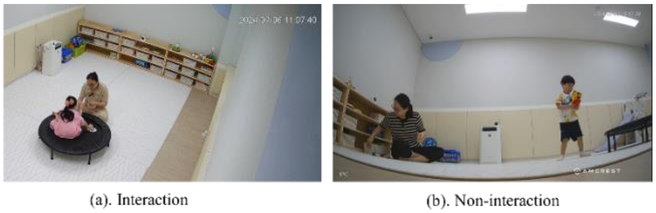

Select parent-child pair for the activity
Participants are in the CCTV room for recording
Start action (t_{begin})
Stop at the specified timestamp (t_{end})
All recordings are completed and saved

Figure 2: Overview of the flowchart of data collection

In this experiment, a pair of parent-child participants was invited to enter the room, and the interactions began shortly after. The data behaviours emphasized the child as the main player while their parent was a cooperator.

1. *Dataset Labelling*

This section describes how we define interaction status. In this study, activities involving face or orienting toward each other, physical contact, gestures, or eye contact inferred through head rotation were segmented and labelled as "Interactions." In contrast, if either participant was engaged in an individual activity without making any of the above behaviours toward the other, the event was labelled as a "Non-interaction". The segmentation process produced 126 segments, categorized into two groups: 83 sub-video files of interaction and 43 sub-video files of non-interaction moments. In the Fig. 3 illustrated the dataset characteristics as interaction and non-interaction activities.



(a). Interaction

(b). Non-interaction

Figure 3: Dataset Characteristics

2. *Data Division*

Data splitting is a key step in machine learning to ensure robust model evaluation. In this study, the dataset was split into two main subsets: 70% for training and 30% for testing. To further enhance the reliability of the model evaluation, 30% of the training set was allocated for cross-validation. This approach helps prevent overfitting and ensures the model can generalize well to unseen data by allowing validation during the training process.

C. *Skeleton Joints Extraction and Object Tracking*

In this study, we utilize the general architecture of YOLO11 and YOLOv8 models, which consists of a Backbone, Neck, and Head [10, 13], responsible for extracting human skeleton joints. We utilize this algorithm since it is capable of achieving an optimal balance between speed, accuracy, and model size, ensuring efficient performance on large-scale data. It is also capable of effectively handling small targets and complex backgrounds [11].

The pre-trained YOLOv8 is responsible for detecting and classifying parents and child in the video sequence. We fine-tuned a pre-trained YOLOv8 model using a combination of public datasets [16, 14] and our dataset, which was labelled as parent, child, and head as the main target objects, using the basic structure of YOLOv8 [15]. This step allows the system to differentiate between parent and child subjects and extract their features. Among the state-of-the-art object detectors, the YOLOv8 model is known for its speed and accuracy. It incorporates several enhancements and improves upon previous YOLO versions.

Human daily activities often originate from specific body points, guiding overall body movement. Therefore, extracting skeleton joints on various body parts is essential for analysing human motion. Head and body movements were identified and continuously tracked over time. In the skeleton joints extraction section, the process begins with the detection of the parent and child's bodies. Bounding box coordinates and identity labels (Parent, Child) are generated around detected persons by the model algorithm and are used to track their movements over time. Next, the detected individuals are continuously tracked across various orientations using the DeepSORT algorithm, which maintains consistent identities throughout the sequence. Lastly, extract the default 17 body joints in 2D space using YOLOv11m-pose [8].

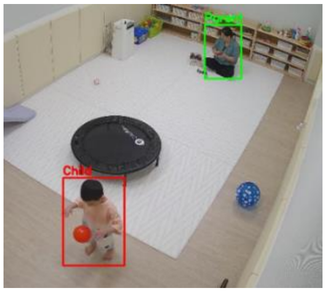


Figure 4: Parent and child classification task

D. Head Detection and Head Rotation Estimation

In the head detection stage, an adapted pre-trained YOLOv8 model was used to detect heads in various orientations. After head detection is completed, the head rotation estimation stage is performed. For this task, we utilized a pre-trained Resnet18 model, originally trained on a public dataset [16] containing head rotation information. This model was then adapted to estimate head orientation on our dataset.

The adapted pre-trained ResNet18 estimates head rotation in terms of pitch, roll, and yaw, and also predicts the head's position in 3D space (x, y, z) following the structure of six degrees of freedom (6DoF-HPE) [17]. The system architecture of head rotation estimation is illustrated in Fig. 4.

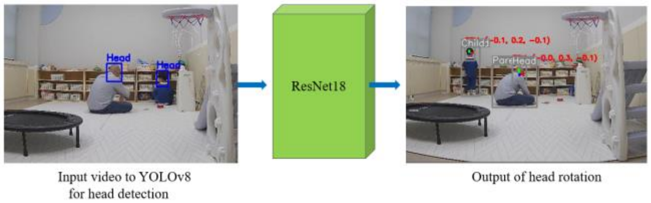


Figure 5: Overview of framework for head rotation estimation

In time-series data analysis, missing values can compromise the integrity of analyses and the performance of predictive models. To address this, we implemented a data preprocessing algorithm that utilizes linear interpolation in the forwarding direction method to estimate and fill the missed data, ensuring the continuity and reliability of the dataset. Initially, the algorithm identifies and replaces missing or zero entries with NaN (Not a Number) to facilitate the interpolation process. The NaN values were then handled using forward-fill (propagating the last known value forward) and backward-fill (propagating the next known value backward) methods to fill the missed values.

This comprehensive approach ensures that all missing entries are appropriately estimated, maintaining the dataset's continuity [21, 22].

Since each sub-video has a different length, the 50% overlap sliding window technique is further utilized to generate fixed-size sub-sequences from continuous time-series data as required by neural networks. We analysed window sizes of 5s using the Random Forest (RF) model.

The core of the sequence generation lies in the sliding window technique. A window of pre-defined length (e.g., 75 time steps or 5 seconds) moves sequentially across the dataset. At each step, a windowed segment of the feature data is extracted as an input sequence, and the label corresponding to the time step immediately after the window is used as the target output. This process transforms the continuous time series into a structured set of samples, where each sample represents a segment of motion over time. The resulting sequences are then reshaped into the format (samples, time steps, features) required for the neural network model.

E. Feature Extraction

The proposed CNN-based feature extraction method enhances temporal data representation using residual connections and Squeeze-and-Excitation (SE) blocks. The model processes sequential input through multiple convolutional and pooling layers. It begins with a Conv1D layer (512 filters, kernel size 7) to capture long-range dependencies, followed by batch normalization, max pooling, and dropout (0.3) for stability and regularization. A residual block with two Conv1D layers (512 filters, kernel size 5) helps retain information and improve gradient flow.

Subsequent layers refine features using 256 and 128 filters, with SE blocks dynamically recalibrating channel importance via global average pooling and attention weighting. The later convolutional layers (128 and 64 filters) incorporate dropout (0.5) and max pooling to prevent overfitting and downsample features. Finally, Global Average Pooling (GAP) and Global Max Pooling (GMP) are combined to capture both broad patterns and dominant features. This hybrid approach ensures a robust feature representation, making the model suitable for action recognition and activity classification.

F. Classifiers

In this study, several machine learning (ML) classifiers were implemented to evaluate their performance in classifying interactions from the dataset. The fusion of extracted features of the proposed method were converted into a feature vector and then passed to seven popular ML classifiers, including Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN), and CatBoost [19], which were used to learn and predict the interaction classes based on extracted features. These classifiers employ various learning methods to capture patterns within the data, each with its strengths and weaknesses.

Different ML classifiers employed various learning techniques to distinguish between two classes: interaction and non-interaction in our dataset, leading to improved performance outcomes. However, in this study, basic parameters were used without fine-tuning to evaluate the classifiers' performance under standard conditions.

III. RESULTS

A. *Experimental Results and Discussion*

An experiment was conducted using the proposed framework on our dataset. Multiple classifiers were evaluated based on various performance metrics, such as precision, recall, F1 score, and accuracy. The best performance results of machine learning classifiers, along with their classification outcomes and confusion matrices, are discussed in this section.

The feature vectors were passed into the validation stage through several machine learning classifiers presented in the section above to evaluate their ability to classify parent-child play activities. Each model was evaluated using stratified 5-fold cross-validation with the data validation set.

Then further tested on the testing set (unseen data) to evaluate classification performance. The average performance metrics across all folds of the cross-validation approach are summarized in Table 1, while model classification performance is reported in Table 2.

Table 1: Average stratifies 5-fold cross-validation performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
RF	93.40 ± 4.22	91.40 ± 5.43	100.00 ± 0.00	95.20 ± 3.06
SVM	85.20 ± 8.01	89.80 ± 6.68	87.40 ± 8.80	88.40 ± 6.41
MLP	74.60 ± 8.87	97.80 ± 4.40	63.20 ± 11.62	76.60 ± 9.95
KNN	84.40 ± 3.56	90.20 ± 6.18	86.40 ± 4.45	88.00 ± 2.68
CatBoost	88.20 ± 3.35	87.40 ± 5.27	96.60 ± 4.98	91.60 ± 2.07

In Table 1, The Random Forest model performed best with the highest accuracy (93.40%) and perfect recall (100%), making it the most reliable for detecting interactions. CatBoost followed closely with strong recall (96.60%) and F1 score (91.60%). KNN and SVM showed balanced results, both around 85% accuracy, but slightly lower recall than CatBoost. MLP had the lowest accuracy (74.60%) and recall (63.20%), indicating it often missed interaction cases despite high precision. Overall, RF is the top choice, with CatBoost as a strong alternative on the data validation set.

Table 2: Results of the proposed method on the testing set

ML Classifiers	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
RF	95.13	94.74	94.62	94.74
SVM	88.46	86.84	87.10	86.84
MLP	88.03	81.58	82.03	81.58
KNN	90.30	89.47	89.63	89.47
CatBoost	92.09	92.11	92.02	92.11

Table 2 presents the classification performance of various machine learning classifiers on the binary classification task of identifying interaction or non-interaction, performance on unseen data, showing that the RF model performs best among the tested classifiers. With a precision of 95.13%, recall of 94.74%, and an F1 score of 94.62%, RF demonstrates a strong ability to correctly identify interactions while minimizing both false positives and false negatives. This indicates that the model is highly reliable and consistent in distinguishing between interactive and non-interactive behavior. CatBoost follows closely with similar performance (92% for precision and recall), making it a solid alternative for efficient deployment. KNN also performs well with an F1 score of 89.63%, showing a good balance. However, both SVM and MLP perform weaker, with SVM achieving an F1 score of 87.10% and MLP underperforming at 82.03%, mainly due to lower recall.

Overall, RF is the most reliable, with CatBoost as a close second.

IV. Conclusions

This study proposes a framework for two-person interaction classification, developed and experimentally evaluated. The effective approach for parent-child interaction recognition is based on the proposed framework, was evaluated with popular machine learning classifiers. The 1D-CNN-based approach provides great overall performance across a wider range of classifiers.

One limitation of this study is that the proposed method is designed for non-verbal settings. However, real-world environments often involve occlusions, overlapping bodies, or motion blur. These factors can reduce classification performance.

However, our proposed method is possible for learning the difference between objects and actions. Additionally, our work supports the design of interactive educational tools and human-computer interaction systems aimed at promoting positive parent-child engagements, which could also be highly useful for clinical and research purposes, given the importance of parent-child relationship quality for child outcomes and mental well-being over the lifespan.

As potential future work, integrating verbal cues (e.g., speech, sound localization) could enhance the robustness of interaction recognition.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No RS-2023-00218176).

References

[1] Horowitz-Kraus, T.; Gashri, C. Multimodal Approach for Characterizing the Quality of Parent–Child Interaction: A Single Synchronization Source May Not Tell the Whole Story. Biology 2023, 12, 241. https://doi.org/10.3390/biology12020241

[2] Lanjekar PD, Joshi SH, Lanjekar PD, Wagh V. The Effect of Parenting and the Parent-Child Relationship on a Child's Cognitive Development: A Literature Review. Cureus. 2022 Oct 22;14(10):e30574. doi: 10.7759/cureus.30574. PMID: 36420245; PMCID: PMC9678477.

[3] Puchała, S.; Kasprzak, W.; Piwowarski, P. Human Interaction Classification in Sliding Video Windows Using Skeleton Data

Tracking and Feature Extraction. Sensors 2023, 23, 6279. https://doi.org/10.3390/s23146279

[4] Tasnim, N.; Islam, M.M.; Baek, J.-H. Deep Learning-Based Action Recognition Using 3D Skeleton Joints Information. Inventions 2020, 5, 49. https://doi.org/10.3390/inventions5030049

[5] Khean, Vesal & Kim, Chomyong & Ryu, Sunjoo & Ahmad, Awais & Hong, Min & Kim, Eun & Kim, Joungmin & Nam, Yunyoung. (2024). Human Interaction Recognition in Surveillance Videos Using Hybrid Deep Learning and Machine Learning Models. Computers, Materials & Continua. 1-10. https://doi.org/10.32604/cmc.2024.056767

[6] A. Senthil Selvi, P. Sibi Aadesh, B. Manoharan and S. Hari Narayanan, "Real-Time Multiple Object Tracking and Object Detection using YOLO v7 and FairMOT Algorithm," 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2023, pp. 1-5, DOI: 10.1109/ICSES60034.2023.10465490

[7] J. -Y. Kim et al., "Analysis of Near-Fall Detection Method Utilizing Dynamic Motion Images and Transfer Learning," in IEEE Access, vol. 13, pp. 26398-26410, 2025, doi: 10.1109/ACCESS.2025.3539449

[8] https://docs.ultralytics.com/tasks/pose/?utm_source=chatgpt.com

[9] https://arxiv.org/pdf/1903.07288

[10] Mao, M.; Hong, M. YOLO Object Detection for Real-Time Fabric Defect Inspection in the Textile Industry: A Review of YOLOv1 to YOLOv11. *Sensors* **2025**, *25*, 2270. https://doi.org/10.3390/s25072270

[11] Shi, Jiayou, et al. "Multi-Crop Navigation Line Extraction Based on Improved YOLO-v8 and Threshold-DBSCAN under Complex Agricultural Environments." Agriculture 14.1 (2023): 45.

[12] Lan, Weishuo & Xu, Jian & Chen, Heyao & Xiao, Xingpeng & Zhao, Mengyuan & Liu, Bo. (2025). Gesture Object Detection and Recognition Based on YOLOv11.

[13] A. Elaoua, M. Nadour, L. Cherroun and A. Elasri, "Real-Time People Counting System using YOLOv8 Object Detection," 2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM), Medea, Algeria, 2023, pp. 1-5, doi: 10.1109/IC2EM59347.2023.10419684.

[14] https://universe.roboflow.com/aniruddha-jmp5a/child-adult-detection-u81pm

[15] arXiv:2310.09492 [cs.CV] (or arXiv:2310.09492v1 [cs.CV] for this version) https://doi.org/10.48550/arXiv.2310.09492

[16] https://universe.roboflow.com/carlos-alberto-castro-zuleta-cnopi/dataset-human-head

[17] Redhwan Algabri, Hyunsoo Shin, Sungon Lee, "Real-time 6DoF full-range markerless head pose estimation, Expert Systems with Applications", https://doi.org/10.1016/j.eswa.2023.122293.

[18] Imran Ahmed, Misbah Ahmad, Joel J.P.C. Rodrigues, Gwanggil Jeon, Sadia Din, A deep learning-based social distance monitoring framework for COVID-19,https://doi.org/10.1016/j.scs.2020.102571.

[19] M. C. Untoro,M. Praseptiawan, M.Widianingsih, I. F. Ashari, andA.Afriansyah, "Evaluation of decision tree, k-NN, Naive Bayes and SVM with MWMOTE on UCI dataset," J. Phys.: Conf. Series, vol. 1477, no.3, 2020, Art. no. 032005. doi: 10.1088/1742-6596/1477/3/032005.

[20] Cigdem Beyan, Alessandro Vinciarelli, and Alessio Del Bue. 2023. Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey. ACM Comput. Surv. 56, 5, Article 109 (November 2023), 41 pages. https://doi.org/10.1145/3626516

[21] Yuhai, O.; Choi, A.; Cho, Y.; Kim, H.; Mun, J.H. Deep-Learning-Based Recovery of Missing Optical Marker Trajectories in 3D Motion Capture Systems. Bioengineering 2024, 11, 560. https://doi.org/10.3390/bioengineering11060560.

[22] Guo, Y., Li, B., Li, Y. et al. Application of a linear interpolation algorithm in radiation therapy dosimetry for 3D dose point acquisition. Sci Rep 13, 4539 (2023). https://doi.org/10.1038/s41598-023-31562-3

Nonverbal Classification of Parent-Child Play Activities Using Skeleton with CNN Extractor and Machine Learning Models

T. Sokea[1], K. Chomyong[2], V. Sreypov[1], N. Yunyoung[3*]

[1]Department of ICT Convergence, Soonchunhyang University, Asan 31538, Republic of Korea
[2]ICT Convergence Research Center, Soonchunhyang University, Asan 31538, Republic of Korea
[3]Emotional and Intelligent Child Care Convergence Center, Soonchunhyang University, Asan 31538, Republic of Korea
**Corresponding Author: Yunyoung Nam. Email: ynam@sch.ac.kr*

Abstract— **Parent-child interactions during play are essential for understanding child development, behavior, and emotional growth. This study presents a nonverbal-based framework for recognizing parent-child play activities using time-series skeleton data extracted from CCTV footage. The proposed system operates by focusing on physical movement and body posture estimation. The proposed framework comprises multi-view video recorded from multi-channel cameras, enabling the model to learn from different angles and make recognition decisions effectively. The YOLOv11 model was used for skeleton joints extraction, while pre-trained YOLOv8 responded to parent-child classification. DeepSORT was simultaneously incorporated for real-time tracking of each detected subject, maintaining identity consistency across frames. Our approach extracts meaningful features from skeletal information using two neural networks involving 1-Dimensional Convolutional Neural Networks (1D-CNN), and Long Short-Term Memory networks (LSTM), and several machine learning classifiers for multi-class classification. The analytical experiments revealed that the 1D-CNN-based method outperformed in classification accuracy, precision, recall, and F1-score. The results show that the proposed framework achieves a high performance and robustness in classifying play activity.**

I. Introduction

Object detection and classification in CCTV footage are essential for analyzing human activities in various environments, including parent-child interaction settings. Parent-child interactions are foundational to children's social, emotional, and cognitive development. Accurate and automatic recognition of these interactions in naturalistic settings can offer significant benefits in developmental psychology, early childhood education, and health monitoring. Traditional methods of observing and evaluating such interactions often rely on manual coding, which is time-consuming and subjective [1]. In recent years, computer vision-based approaches have been explored to automate the detection and classification of human interactions [2], with a growing interest in applications involving parent-child behavior analysis [3].

Human pose estimation and skeleton-based action recognition have proven to be effective in capturing subtle interactions between individuals. Models such as OpenPose [4, 14], PoseNet [5], and more recent YOLO-based pose estimators [6, 7] have enabled robust keypoint detection in challenging environments. However, the presence of multiple closely interacting individuals, such as a parent and a child in close proximity, introduces challenges in accurately distinguishing keypoints and preserving person identity over time [8].

To address these issues, various approaches have been proposed. Tracking algorithms such as Deep SORT [9] and BoT-SORT [10] have improved the temporal consistency of detections, but they often struggle with ID switching when people are close together or partially occluded. Moreover, keypoint extraction models aim to produce overlapping or inconsistent joint estimations when individuals are in physical contact or appear similar in size and motion [11]. These limitations pose a significant challenge in accurately modeling parent-child interactions. The authors of [27] highlighted that YOLOv11 outperforms all known object detectors in both speed and accuracy. YOLOv11m-pose offered by YOLO11 pose models, designed specifically for human pose estimation, not just object detection. This model treats object detection as a regression problem that directly predicts bounding boxes and class labels. Human daily activities often originate from specific body points, guiding overall body movement. Therefore, extracting skeleton joints on various body parts is essential for analyzing and understanding human motion and interaction by mapping body joints into a structure representation.

In this work, we present a study focused on nonverbal behaviors. Nonverbal cues such as body gestures, facial expression, motion, eye contact, or any physical actions serve as important indicators of parent-child interaction styles, providing a deeper understanding of their relational dynamics [12, 13]. In this study, we aim to automatically recognize nonverbal parent-child play activity from a video sequence, identify play activities as individual subjects based on skeletal information, and classify the play activity into four classes: play ball, playing with toys, jumping, and playing on the slide. Recognizing these activities can assist in the early detection of developmental delays and support interventions in early childhood education or behaviour monitoring.

II. Method

In this section, we describe the dataset collection process and outline the proposed methodology for recognizing parent-child play activities. The process begins with data preprocessing, which includes human detection and pose estimation. Following preprocessing, three main approaches were explored for identifying parent and child activities: (1) human classification using a pre-trained YOLOv8 model to distinguish between parent and child, (2) pose estimation to extract body skeleton joints for each individual, and (3) interpolation techniques to smoother the extracted data. After preprocessing, feature extraction was performed using two different methods: CNN, and LSTM, allowing for comparison to determine the most effective feature extractor. Subsequently, popular classification models were employed to classify the play activities. Finally, cross-validation was used to evaluate model performance. All computations were performed using Python 3.8, Ultralytics YOLO v8.3.74, PyTorch 2.4.1, OpenCV 4.11, Visual Studio Code, and DeepSORT v1.3.2.

1) Dataset

Our video surveillance contains many activities (free play) in pairs of participants. The dataset consists of 26 parent-child pairs, where the children in 2-6 years old engaged in a child playing environment (indoor environment). The dataset aimed at capturing natural play activities between a parent and a child with their interest and enjoyment.

Two RGB cameras (labelled Camera1 and Camera4) were strategically placed to capture interactions from two different angles for video data collection. All channel cameras were positioned for top-view footage as presented in Figure 1. The collected videos range from 10 to 15 minutes in length, captured from different angles of the same scene, at a resolution of 3840 x 2160 pixels, 15 fps (frames per second). The detailed data collection procedure is shown in Fig. 2.

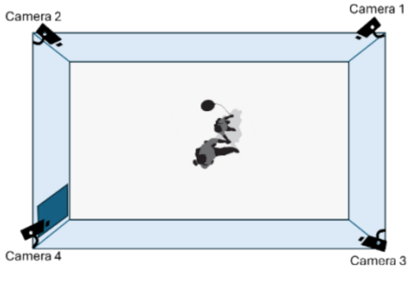


Fig 1: Room environment and four camera positions on the top views.

2) Proposed Methodology

The proposed model in the Figure 3 is consists of four main components: pre-trained models, tracking, preprocessing, and deep learning. An overview of the method is shown in Fig. 2. We utilize two pre-trained YOLO models:

1. Parent-child classification model: A pre-trained YOLOv8-based [26] model uses to classify individuals as either a parent or a child.

2. Keypoints estimation model: Extract skeletal information from individuals using yolo11m-pose [28], which is represented the human motions.

For objects tracking, DeepSORT [39, 40] uses to continuously track individual parent and child.

Once the play activities were collected by body joint as the skeleton data, all relevant information is stored in a CSV file, compiling data from each sub-video.

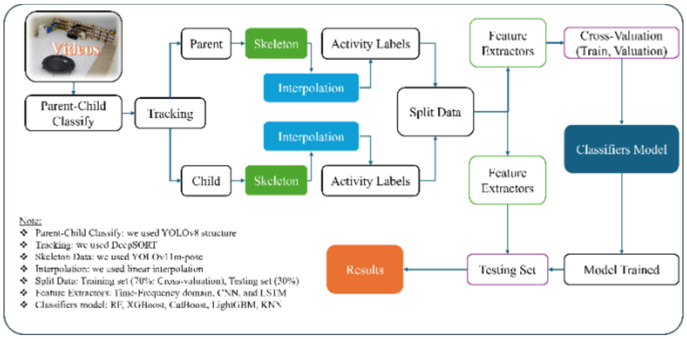

Videos
Parent-Child Classify
Tracking
Parent
Skeleton
Interpolation
Activity Labels
Child
Skeleton
Interpolation
Activity Labels
Split Data
Feature Extractors
Feature Extractors
Cross-Valuation (Train, Valuation)
Classifiers Model
Model Trained
Testing Set
Results
Note:
Parent-Child Classify: we used YOLOv8 structure
Tracking: we used DeepSORT
Skeleton Data: we used YOLOv11m-pose
Interpolation: we used linear interpolation
Split Data: Training set (70%: Cross-valuation), Testing set (30%)
Feature Extractors: Time-Frequency domain, CNN, and LSTM
Classifiers model: RF, XGBoost, CatBoost, LightGBM, KNN

Fig 2: Overview of the proposed model identifying interaction and non-interaction

3) Data Preprocessing

The data preprocessing stage is necessary for preparing raw data into a suitable format for modelling and classification. It includes organizing, structuring, and augmenting, etc., to allow the machine learning and deep learning to understand the pattern. The video was read across frames using the OpenCV (Open-Source Computer Vision) library and resized from the original size to 960 × 960 using the YOLO library. High-resolution frames at the original size (3840 x 2160) contain a lot of pixels, which increases the computational and memory usage. Resizing frame resolution helps to standardize model input, enhance processing speed, and reduce memory usage.

A. Data Labeling

In this study, our research specifically aimed to classify play activity via playing behaviors into four different categories, including jumping, playing with the ball, playing on the slides, and playing with toys. We categorized the classes based on an activity and an object the child and parent interact with, or the specific behaviors observed in the video. Those are defined by: “Jumping” class. When the subject was jumping on a mini trampoline by bending the knees, pushing down, and bouncing up into the air repeatedly, the action was categorized as jumping. “Play toys” class, when the subject is playing with toys using hands to pick up, hold, move, or manipulate toys, we categorized them as play toys. The “Play ball” class refers to the activity where the subject plays with a ball, showing behaviours such as holding, throwing, rolling, bouncing, or kicking it, either alone or with others. The activity where the subject interacts with a slide, showing behaviors such as climbing up to the top of the slide and sliding back down, we labelled as “Play on the slide”.

B. Training

Data splitting is a key step in machine learning to ensure robust model evaluation. In this study, the dataset, composed of two combined channel cameras, was split into two main subsets: 70% for training and 30% for testing. To further enhance the reliability of the model evaluation, 30% of the training set was allocated for cross-validation. This approach helps prevent overfitting and ensures the model can generalize well to unseen data by allowing validation during the training process.

C. *Parent-child classification model*

To effectively classify parent and child, we fine-tuned a pre-trained YOLOv8 model using a combination of public datasets [25] and our dataset, which was labelled as parent and child as the main target objects, using the basic structure of the YOLOv8 model [26]. The adapted pre-trained YOLOv8 classifies individuals in playtime videos, determining them as either parent or child, as shown in Figure 3.

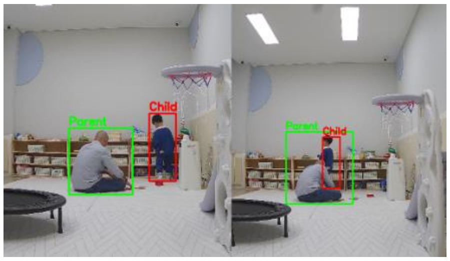


Fig 3: Example of Parent and child classification task

After classification, DeepSORT tracking was applied to assign and maintain consistent IDs for the classified parents and children across frames. This ensured continuous and reliable tracking of each segment throughout the video.

D. *Skeleton Data Extraction*

This section involves skeleton data extraction, correction, and transforming raw data into a suitable format for modeling. In the skeleton data extraction task, we utilized the YOLOv11m-pose model to extract skeleton joints of the individual person in the frames. In the skeleton joints extraction section, person and body movements were identified and continuously tracked over time. We extract the skeletal information of human participants who interacted either with each other or alone in the video sequences.

The performance begins with the detection of the parents' and the child's bodies. Second, continuously tracking the detected bodies in various orientations through the support of DeepSORT. Third, assigning identification (Parent, Child) to the tracked bodies. Lastly, extract the default 17 major body joints data from each detected human body in 2D (x, y) space using YOLOv11m-pose [28]. The model extracts skeleton joints from video with frames by frames for further analysing human actions and understanding activity patterns. The detected keypoints (green) represent anatomical landmarks of both the parent and child, while the skeletal connections (blue) visualize their body posture and movements shown in Figure 4. This approach enhances the accuracy of action recognition systems and contributes to advances the state-of-the-art in human activity analysis.

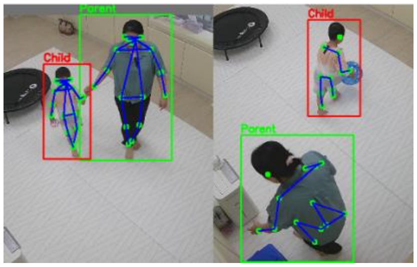


Fig 4: Skeleton joints extraction task

Importantly, since our data collection is operated within a free-play setting, occlusions and any mistakes during the object detection stage will lead to missing raw data or incomplete skeleton data from each frame [29]. In this time-series data analysis, missing data in frames can affect the integrity of analyses and the performance of predictive models. To address this issue, we implemented a linear interpolation algorithm in the forwarding direction to estimate and fill gaps of missed frames in skeleton data, ensuring the continuity and reliability of the dataset [14, 30, 31]. This comprehensive approach ensures that all missing entries are appropriately estimated and filled in, maintaining the dataset's continuity.

Since each sub-video has a different length, the 50% overlap sliding window technique is further utilized to generate fixed-size sub-sequences from continuous time-series data. We analyzed various window sizes, including 5s, 10s, and 15s, to find the most effective segmentation size for activity classification, using the Random Forest (RF) model.

Additionally, we employed a stratified 5-fold cross-validation approach to evaluate the model, averaging the results over five iterations.

4) *Feature extraction*

Feature extraction is essential for improving the learning performance of classification models. In this study, we applied two different feature extraction techniques separately including 1D-CNN and LSTM, to extract meaningful features from segmented skeleton data capturing various parent-child activities.

A. *1D-CNN-Based Feature Extraction*

1D-CNNs are especially suitable for such tasks, as they efficiently model temporal dependencies across features while maintaining low computational complexity [32]. The 1D-CNN architecture is composed of six convolutional blocks, each carefully structured to extract temporal features. The CNN architecture is specifically designed for time-series feature extraction, incorporating hyperparameter tuning to optimize performance. The model begins with an input layer that accepts data shaped by time steps and feature columns. The first convolutional block applies a Conv1D layer with 512 filters, a kernel size of 7, followed by Batch Normalization (BN) and ReLU activation, extracting broad temporal patterns. This is followed by MaxPooling1D to downsample the sequence (pool size = 2) and Dropout (0.3) for regularization.

A residual block is included next, where two Conv1D layers (each with 512 filters and kernel size = 5) are stacked, each followed by BN and ReLU. This residual connection helps retain information across layers and prevents weak learning signals in deeper layers. Another MaxPooling1D layer further downsamples the data, followed by a Conv1D layer with 256 filters and kernel size = 3, then passed through a Squeeze-and-Excitation (SE) block, which adaptively recalibrates channel-wise features. After another MaxPooling1D, the feature map splits into a parallel branch.

This second branch starts with Conv1D (128, k=3) + BN, followed by another SE block and AveragePooling1D (size = 2). It continues through two more Conv1D layers (128 and 64 filters respectively, both with kernel size = 3), each followed by Dropout (0.5) to prevent overfitting. Both paths conclude with Global Average Pooling and Global Max Pooling, which are merged using an Add operation, effectively aggregating spatial information globally from both representations. The merged result is the final feature output, representing an enriched embedding of the time-series input.

This CNN architecture has undergone hyperparameter tuning, as seen in the choice of filter sizes (512, 256, 128, 64), kernel sizes (3, 5, 7), dropout rates (0.3, 0.5), and use of SE blocks and residual connections, all selected to enhance model accuracy, generalization, and computational efficiency in the feature extraction process.

B. LSTM-Based Feature Extraction

The LSTM model is well-suited for sequence modelling tasks due to their ability to retain both short- and long-term temporal relationships [33, 34, 35]. The architecture consists of two stacked LSTM layers with 128 and 64 units, respectively. The first LSTM layer is configured to return sequences, allowing the second LSTM layer to process the full temporal output of the first layer. This hierarchical structure enables the model to learn both short-term and long-term dependencies across sequential keypoints. Following the LSTM layers, a fully connected dense layer with 64 neurons and a ReLU activation function is employed to project the learned temporal features into a compact and discriminative feature space. This LSTM-based extractor effectively models the sequential relationships inherent in human motion, providing rich temporal representations for the classification of parent-child play activities.

These feature vectors are then passed to a traditional machine learning classifier for the prediction task.

5) *Classifiers*

In this study, several machine learning (ML) classifiers were implemented to evaluate their performance in classifying interactions from the dataset. The fusion of extracted features of the proposed method was converted into a feature vector and then passed to five popular ML classifiers, including Random Forest (RF), K-Nearest Neighbors (KNN), XGBoost, and CatBoost, and LightGBM [36, 41], which were used to learn and predict the play activity based on extracted features. These classifiers employ various learning methods to capture patterns within the data, each with its strengths and weaknesses. However, in this study, basic parameters were used without fine-tuning to evaluate the classifiers' performance under standard conditions.

6) *Evaluation*

The performance of each model was assessed using standard evaluation metrics, including accuracy, recall, precision, and F1-score [41]. To ensure robustness and generalizability of the results, we employed stratified 5-fold cross-validation during both the sliding window evaluation and the feature extraction experiments. This approach maintains the original class distribution across folds, providing more reliable performance estimates.

To further evaluate the significance of performance differences, we conducted independent two-sample t-tests. Specifically, independent two-sample t-tests were employed to compare the mean F1 score between varying feature extraction methods (1D-CNN and LSTM). The t-test evaluated whether the observed differences in model performance were statistically significant. The resulting p-values were analyzed with a common threshold of $p < 0.05$. In cases where the p-values are larger than this threshold, it's indicated that the differences were not statistically significant [37].

III. RESULTS AND DISCUSSION

1) *Data Distributions*

The dataset consists of time-series skeleton data representing a variety of parent-child play activities. Each activity segment was derived from continuous video recordings, with sub-videos generated through temporal segmentation during specific actions. Figure 10 presents the distribution across all classes in terms of percentage.

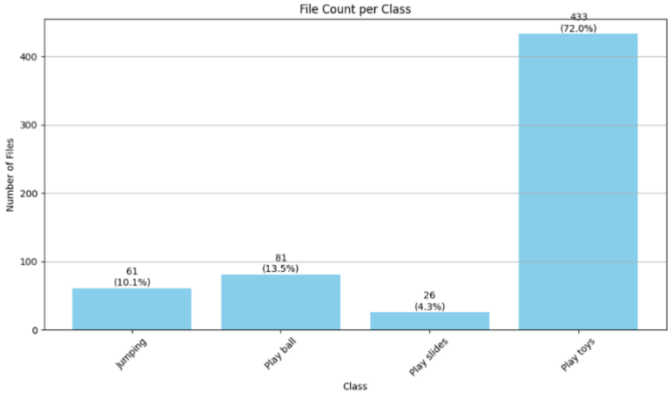


Figure 5: Parent-Child play activity distribution

The combined class distribution highlights a clear class imbalance in the dataset. The Playing toys account for 72.0% of all samples, followed by the Play ball (13.5%). Conversely, the Playing on the slide and Jumping are the least frequent, contributing only 4.3% and 10.1% of the total dataset, respectively.

Hence, we applied a data augmentation strategy using the K-MeansSMOTE [42] oversampling method only during model training to obtain a balanced training set to help improve model performance by allowing the model to learn to recognize more of the minority classes. While the test set was kept as original to reflect a real-world, naturally imbalanced dataset, to evaluate generalization performance.

2) *Analysis*

Table 1 presents the sliding window comparison across varying durations (5s vs 10s, 5s vs 15s). The experiment was conducted using stratified 5-fold cross-validation incorporated with RF model to evaluate its effectiveness on multi-class classification performance for various play activities.

Table 1: Average validation metrics across 5-fold cross-validation

Window Size	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
5 sec	96.20 ± 0.23	88.89 ± 1.97	64.05 ± 1.41	72.06 ± 1.61
10 sec	95.81 ± 0.41	88.54 ± 6.96	61.85 ± 1.68	69.84 ± 2.80
15 sec	95.91 ± 0.53	69.85 ± 1.25	48.24 ± 1.97	52.74 ± 2.95

The 5-second window outperformed the others, achieving the highest F1 score (72.06 ± 1.61), along with strong precision (88.89 ± 1.97) and recall (64.05 ± 1.41). While all window sizes reported similarly high accuracy (around 95%), the considerably lower precision, recall, and F1 scores are attributable to the class imbalance in the dataset.

Furthermore, A statistical paired t-test is conducted to determine whether there is a significant difference in performance between the two groups. The performance between 5s vs 10s window size produced a $p > 0.05$, confirming that there is no statistically significant difference in performance, while the performance between 5s vs 15s window size produced a $p < 0.05$, confirming that there is a statistically significant difference in performance.

However, we selected the 5-second window size based on both practical and theoretical considerations. The main reason is that all video segments in our dataset contain at least 5 seconds in length. Choosing a window larger than 5 seconds would lead to data loss by excluding valid shorter segments, which may result in bias or less generalizable models. In addition, the 5-second window size already achieves a good performance among others. These support the selection of a 5-second window as the most effective window size for classifying the data with diverse play behaviour conditions.

3) *Evaluation of Feature Extraction Methods*

The performance comparison between 1D-CNN and LSTM feature extractors is presented in Table 2, evaluated using RF classifier with 5-fold cross-validation. The average results over five folds revealed that the 1D-CNN outperforms the LSTM across evaluation metrics. 1D-CNN achieved the highest classification accuracy of 96.20 ± 0.23, outperforming the LSTM's accuracy of 94.54 ± 0.18. In terms of F1-score, which balances precision and recall, 1D-CNN also leads with 72.06 ± 1.61 compared to LSTM's 58.38 ± 2.48, indicating that 1D-CNN yields more reliable and consistent predictions.

Table 2: Performance comparison between feature extractors

To further validate the performance difference between feature extractors, a statistical paired t-test was conducted on their classification accuracy scores. The t-test produced a p-value of 0.0001, which is well below the standard significance threshold of 0.05 ($p < 0.05$). This indicates that the difference in accuracy between the two models is statistically significant.

4) *Results*

After identifying that 1D-CNN-based as the most effective feature extractor, we then passed those feature vectors into the validation stage through several machine learning classifiers presented in the section above to evaluate their ability to classify play activities on individual subjects as parent and child. Each model was evaluated using stratified 5-fold cross-validation with the data validation set.

Then further tested on the testing set (unseen data) to evaluate classification performance. Table 3 summarizes the average performance metrics across all folds of the cross-validation, while model classification performance is reported in Table 4.

Table 3: Average stratifies 5-fold cross-validation performance using 1D-CNN-based

Model	Subjects	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
RF	Parent	97.59 ± 0.43	74.27 ± 17.43	71.51 ± 17.28	72.78 ± 17.36
	Child	96.32 ± 0.42	92.45 ± 2.56	74.85 ± 4.69	81.00 ± 3.47
XGBoost	Parent	97.59 ± 0.38	69.69 ± 7.16	78.58 ± 16.65	71.82 ± 10.18
	Child	96.30 ± 0.23	89.78 ± 2.84	75.85 ± 4.44	81.23 ± 3.77
CatBoost	Parent	97.73 ± 0.21	72.38 ± 16.52	67.77 ± 10.03	97.73 ± 0.21
	Child	96.30 ± 0.38	87.73 ± 2.28	79.37 ± 3.90	82.84 ± 2.16
LightGBM	Parent	97.84 ± 0.39	67.99 ± 14.42	66.27 ± 14.05	67.06 ± 14.21
	Child	96.48 ± 0.33	91.18 ± 2.12	77.42 ± 5.15	82.61 ± 3.45
KNN	Parent	97.53 ± 0.20	72.03 ± 12.22	78.35 ± 17.37	72.77 ± 13.40
	Child	96.24 ± 0.40	87.20 ± 1.79	81.40 ± 2.57	83.89 ± 0.93

For parent classification, all models achieved high accuracy (above 97.5%), with LightGBM slightly outperforming others in accuracy (97.84%). However, RF and KNN showed the best F1 scores (~72.8%), indicating better balance between precision and recall. For child classification, KNN achieved the best overall performance, with the highest F1 score (83.89%) and recall (81.40%), while LightGBM had the highest accuracy (96.48%). Although RF had the highest precision (92.45%), its lower recall led to a reduced F1 score.

Overall, KNN demonstrated the most balanced and consistent results on the validation stage.

Table 4: Model performance on the testing set using 1D-CNN-based

Feature Extractor	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1D-CNN	96.20 ± 0.23	88.89 ± 1.97	64.05 ± 1.41	72.06 ± 1.61
LSTM	94.54 ± 0.18	92.53 ± 2.72	49.78 ± 1.69	58.38 ± 2.48

Model	Subjects	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
RF	Parent	0.9736	0.6149	0.5564	0.5816

	Child	0.9608	0.9523	0.7873	0.8458
XGBoost	Parent	0.9736	0.5745	0.5866	0.9736
	Child	0.9614	0.9325	0.8272	0.8728
CatBoost	Parent	0.9757	0.6143	0.5716	0.5909
	Child	0.9631	0.9243	0.8014	0.8497
LightGBM	Parent	0.9757	0.6113	0.5752	0.5918
	Child	0.9648	0.9199	0.8346	0.8346
KNN	Parent	0.9736	0.6051	0.5781	0.5908
	Child	0.9585	0.8724	0.8292	0.8485

The experiment on the test set for both parent and child activity recognition using 1D-CNN features revealed that all models achieved high accuracy (~97.36% to 97.57%) on the parent dataset. CatBoost and LightGBM had the highest accuracy (97.57%), while LightGBM slightly led in F1 score (59.18%). Overall, precision and recall for all models on parent data were relatively low, indicating a performance gap in capturing parent-specific play activities.

In contrast, performance on the child dataset was stronger across all metrics. XGBoost achieved the best F1 score (87.28%) with high precision (93.25%) and recall (82.72%), followed closely by CatBoost and KNN. LightGBM had the highest accuracy (96.48%) but a lower F1 score (83.46%) due to a drop in precision.

Overall, XGBoost performed best for child activity recognition, while LightGBM and CatBoost performed better for parent activity recognition. Their results show that it generalizes well to unseen data.

Although KMeansSMOTE was applied to balance the training set, the precision, recall, and F1-scores remain lower than the accuracy. This is because the test set was left imbalanced to reflect real-world data distribution. In such cases, accuracy can appear high due to correct predictions on the majority class, while the model may still struggle to correctly detect the minority class, lowering the recall and F1-score. These metrics are more sensitive to class imbalance and provide a clearer sign of how well the model handles all classes, especially the minority class.

IV. CONCLUSION

This study presents a nonverbal-based framework for classifying parent-child play activities using time-series skeleton data captured from surveillance video, using deep learning-based feature extraction and traditional machine learning classifiers. Through a systematic evaluation process, we optimized three critical components of the pipeline: accurate object detection, feature extraction, and classifier performance, ensuring the model's effectiveness in real-world conditions.

Overall, the results confirmed that nonverbal interaction analysis based on skeletal movement data can achieve high classification accuracy when incorporated with three critical components mentioned above. The proposed framework is suitable and effective for real-time nonverbal behaviour recognition in child development monitoring, educational interaction analysis, or assistive care applications. By focusing on body posture and movement without depending on verbal cues, the study contributes to the growing field of nonverbal human object interaction (HOI) recognition and human interaction recognition (HIR) in future work, particularly in scenarios where verbal data may be unavailable, intrusive, or culturally inappropriate.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No RS-2023-00218176).

REFERENCES

[1] M. C. Baker, "Parent-child interaction: Methodological advances and new directions," Developmental Psychology, vol. 48, no. 3, pp. 689–702, 2012.

[2] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," International Journal of Wavelets, Multiresolution and Information Processing, vol. 2, no. 02, pp. 121–132, 2004.

[3] A. Borghi et al., "Child-centered action recognition in videos via attention-based spatiotemporal modeling," in Proc. CVPR Workshops, 2020, pp. 2250–2258.

[4] Z. Cao et al., "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," in Proc. CVPR, 2017, pp. 7291–7299.

[5] A. Papandreou et al., "Towards accurate multi-person pose estimation in the wild," in Proc. CVPR, 2017, pp. 4903–4911.

[6] C. Wang et al., "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv:2207.02696, 2022.

[7] G. Jocher et al., "YOLOv8: Ultralytics YOLOv8," [Online]. Available: https://github.com/ultralytics/ultralytics, 2023.

[8] X. Wang et al., "Person re-identification: Past, present and future," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 287–307, Jan. 2022.

[9] N. Wojke et al., "Simple online and realtime tracking with a deep association metric," in Proc. ICIP, 2017, pp. 3645–3649.

[10] Y. A. Bochinski et al., "BoT-SORT: Better tracking by optimizing sort," in Proc. CVPR Workshops, 2022, pp. 3467–3475.

[11] J. Sarandi, et al., "Robust joint tracking in multi-person pose estimation," in Proc. CVPR, 2020, pp. 8075–8084.

[12] Peitong Li, Hui Lu, Ronald Poppe, and Albert Ali Salah. 2023. Automated Detection of Joint Attention and Mutual Gaze in Free Play Parent-Child Interactions. https://doi.org/10.1145/3610661.3616234

[13] Cigdem Beyan, Alessandro Vinciarelli, and Alessio Del Bue. 2023. Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey. ACM Comput. Surv. 56, 5, Article 109 (November 2023), 41 pages. https://doi.org/10.1145/3626516

[14] Puchała, S.; Kasprzak, W.; Piwowarski, P. Human Interaction Classification in Sliding Video Windows Using Skeleton Data Tracking and Feature Extraction. Sensors 2023, 23, 6279. https://doi.org/10.3390/s23146279

[15] Khean, Vesal & Kim, Chomyong & Ryu, Sunjoo & Ahmad, Awais & Hong, Min & Kim, Eun & Kim, Joungmin & Nam, Yunyoung. (2024). Human Interaction Recognition in Surveillance Videos Using Hybrid Deep Learning and Machine Learning Models. Computers, Materials & Continua. 1-10. https://doi.org/10.32604/cmc.2024.056767

[16] A. Senthil Selvi, P. Sibi Aadesh, B. Manoharan and S. Hari Narayanan, "Real-Time Multiple Object Tracking and Object Detection using YOLO v7 and FairMOT Algorithm," DOI:10.1109/ICSES60034.2023.10465490

[17] Wang L, Su B, Liu Q, Gao R, Zhang J, Wang G. Human Action Recognition Based on Skeleton Information and Multi-Feature Fusion. Electronics. 2023; 12(17):3702. https://doi.org/10.3390/electronics12173702

[18] Hendra, Jaya (2022) Object Recognition Applications in the Home for Early Children's Education Based on MobileNet. International Journal of Scientific Engineering and Science, 6 (6). pp. 7-12. ISSN 2456-7361 URI: http://eprints.unm.ac.id/id/eprint/30363.

[19] Puchała, S.; Kasprzak, W.; Piwowarski, P. Human Interaction Classification in Sliding Video Windows Using Skeleton Data Tracking

and Feature Extraction. Sensors 2023, 23, 6279. https://doi.org/10.3390/s23146279.

[20] B. Karaca, A. A. Salah, J. Denissen, R. Poppe and S. M. C. de Zwarte, "Survey of Automated Methods for Nonverbal Behavior Analysis in Parent-Child Interactions," 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), Istanbul, Turkiye, 2024, pp. 1-11, doi: 10.1109/FG59268.2024.10582009.

[21] Mao, M.; Hong, M. YOLO Object Detection for Real-Time Fabric Defect Inspection in the Textile Industry: A Review of YOLOv1 to YOLOv11. Sensors 2025, 25, 2270. https://doi.org/10.3390/s25072270

[22] Shi, Jiayou, et al. "Multi-Crop Navigation Line Extraction Based on Improved YOLO-v8 and Threshold-DBSCAN under Complex Agricultural Environments." Agriculture 14.1 (2023): 45.

[23] M. Alruwaili et al.: Deep Learning-Based YOLO Models for the Detection of People with Disabilities. 2024.

[24] A. Alaoui, M. Nadour, L. Cherroun and A. Elasri, "Real-Time People Counting System using YOLOv8 Object Detection," 2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM), Medea, Algeria, 2023, pp. 1-5, doi: 10.1109/IC2EM59347.2023.10419684.

[25] https://universe.roboflow.com/aniruddha-jmp5a/child-adult-detection-u81pm

[26] https://docs.ultralytics.com/models/yolov8/

[27] Lan, Weishuo & Xu, Jian & Chen, Heyao & Xiao, Xingpeng & Zhao, Mengyuan & Liu, Bo. (2025). Gesture Object Detection and Recognition Based on YOLOv11.

[28] https://docs.ultralytics.com/tasks/pose/

[29] Peitong Li, Hui Lu, Ronald Poppe, and Albert Ali Salah. 2023. Automated Detection of Joint Attention and Mutual Gaze in Free Play Parent-Child Interactions. https://doi.org/10.1145/3610661.3616234

[30] Yuhai, O.; Choi, A.; Cho, Y.; Kim, H.; Mun, J.H. Deep-Learning-Based Recovery of Missing Optical Marker Trajectories in 3D Motion Capture Systems. Bioengineering 2024, 11, 560. https://doi.org/10.3390/bioengineering11060560.

[31] Guo, Y., Li, B., Li, Y. et al. Application of a linear interpolation algorithm in radiation therapy dosimetry for 3D dose point acquisition. Sci Rep 13, 4539 (2023). https://doi.org/10.1038/s41598-023-31562-3

[32] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, Daniel J. Inman, 1D convolutional neural networks and applications: A survey, Mechanical Systems and Signal Processing, Volume 151, 2021, 107398, ISSN 0888-3270, https://doi.org/10.1016/j.ymssp.2020.107398

[33] Y. Yu, X. Si, C. Hu and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," in Neural Computation, vol. 31, no. 7, pp. 1235-1270, July 2019, doi: 10.1162/neco_a_01199.

[34] Duan, Ziheng, et al. "Multivariate time-series classification with hierarchical variational graph pooling." Neural Networks 154 (2022): 481-490.

[35] Zhang, S.; Li, Y.; Zhang, S.; Shahabi, F.; Xia, S.; Deng, Y.; Alshurafa, N. Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. Sensors 2022, 22, 1476. https://doi.org/10.3390/s22041476

[36] M. C. Untoro,M. Praseptiawan, M.Widianingsih, I. F. Ashari, andA.Afriansyah, “Evaluation of decision tree, k-NN, Naive Bayes and SVM with MWMOTE on UCI dataset,” J. Phys.: Conf. Series, vol. 1477, no.3, 2020, Art. no. 032005. doi: 10.1088/1742-6596/1477/3/032005.

[37] Teng, S.; Kim, J.-Y.; Jeon, S.; Gil, H.-W.; Lyu, J.; Chung, E.H.; Kim, K.S.; Nam, Y. Analyzing Optimal Wearable Motion Sensor Placement for Accurate Classification of Fall Directions. Sensors 2024, 24, 6432. https://doi.org/10.3390/s24196432

[38] Shyaa, Tahreer & Hashim, Ahmed. (2024). Enhancing real human detection and people counting using YOLOv8. BIO Web of Conferences. 97. 00061. 10.1051/bioconf/20249700061.

[39] Narinder Singh Punn, Sanjay Kumar Sonbhadra, Sonali Agarwal, Gaurav Rai. Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques. https://doi.org/10.48550/arXiv.2005.01385

[40] R., Athilakshmi & Sainagakrishna, Pulavarthi & Kota, Sreya & Muddangula, Chandra Kiran Teja & Venkatesh, Tummala & Prasad, V. (2023). Enhancing Real-Time Human Tracking using YOLONAS-DeepSort Fusion Models. DOI:10.1109/ICECA58529.2023.10394864

[41] Mohammed, Mohammed & Kadhem, Suhad & Maisa, & Ali, A. (2021). Insider Attacker Detection Using Light Gradient Boosting Machine. 1. 48-66.

[42] Felix Last, Georgios Douzas, Fernando Bacao. (2017). Oversampling for Imbalanced Learning Based on K-Means and SMOTE. https://doi.org/10.48550/arXiv.1711.00837

Tangible Interaction-Based Game Bridging the Physical and Virtual Worlds

Jung-Hun Ryu[1], Ji-Won Choi[1], Soo-Hyun Lim[1], Hae-Vin Lee[1], Won-Seop Shin[2], Yong-Hoon Jung[2] and Sang-Hyun Seo[1,*]
[1]*School of Art and Technology, Chung-Ang University, Anseong-si 17546, South Korea*
[2]*Department of Applied Art and Technology, Chung-Ang University, Anseong-si 17546, South Korea*
*Contact: sanghyun@cau.ac.kr, phone +82-10 7273 0318

Abstract— **This paper introduces a tangible interaction-based game designed to enhance player immersion and engagement. The game adopting a puzzle format combines physical block manipulation with a virtual environment, offering a more intuitive interaction method. To overcome traditional game's limitation that rely solely on visual and auditory stimuli, this system let players manipulate physical blocks. The players' actions are immediately reflected in the positions and states of digital characters through an Arduino-powered interface synchronized with Unreal Engine. This approach delivers a more immersive and accessible puzzle-solving experience. Future research will focus on evaluating user responses and assessing the system's effectiveness in enhancing usability and cognitive engagement.**

I. Introduction

Continuous advancements in computer technology have significantly impacted our daily lives. In recent years, it is difficult to imagine life without computers. Among the numerous fields influenced by this trend, video games have emerged as a central component of contemporary popular culture. Despite negative perceptions, such as concerns about violence or addiction, video games are increasingly recognized as a distinct and evolving form of art.

A defining characteristic of video games is their real-time interactivity. Unlike traditional artistic media, video games provide immediate feedback based on player inputs, reinforcing user immersion through iterative interactions [1].

However, current video game experiences predominantly rely on visual and auditory stimuli. The digital reproduction of tactile, olfactory, and gustatory sensations remains technologically challenging, presenting inherent limitations to achieving deeper, multisensory immersion in virtual worlds.

To address this limitation, this study introduces tangible interaction, which employs physically touchable objects to simulate real-world experiences within the game. This approach aims to bridge the gap between the physical and virtual worlds, offering users a more immersive and accessible gaming experience.

II. Game Design

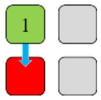
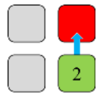
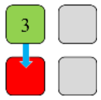
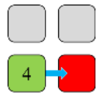
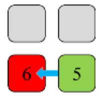
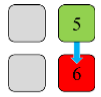
A. Puzzle Design

Our goal is to effectively implement tangible interaction; thus, complex game formats are unnecessary. We selected a puzzle format due to its inherent simplicity, consisting of basic components such as pieces, objectives, and rules [2].

In our game, six in-game characters residing in individual rooms act as the puzzle pieces. Each character exists in either a 'satisfied' or 'unsatisfied' state and has the ability to influence others. Table 1 summarizes the influence exerted by each character. Depending on this influence, the state of each character is determined to be either satisfied or unsatisfied. In the example column, the character depicted in green influences another character depicted in red, resulting in the red character's unsatisfied state. This logic is disclosed within the game, inducing the player to discover and utilize it to achieve the game's objective by manipulating the characters.

The objective is for players to arrange the characters so that all characters reach the satisfied state. The game follows a simple rule: players can swap the positions of two blocks.

TABLE I
Character's Influence Chart

Character	Influenced Target Position	Example
[1]	one floor below	
[2]	one floor above	
[3]	one floor below	
[4]	right room	
[5]	[6] who is on the same floor, the floor above, or the floor below	
[6]	None	

B. Components for Tangible Interaction

As shown in Fig. 1, the game consists of three main hardware components: a display, blocks, and a case.

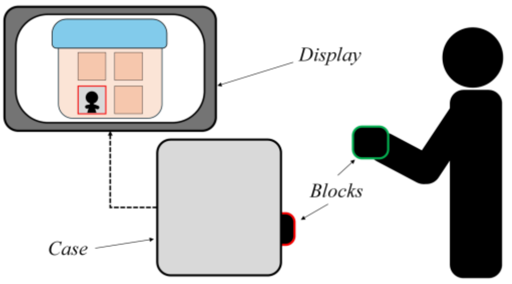


Fig. 1 Components of the Puzzle

Rendered by a game engine, the display dynamically updates the virtual environment in response to players' block arrangements. Physical blocks, each representing a character, were chosen for their simplicity and intuitiveness. This design allows players to engage with the game with minimal instruction—simply by placing the blocks into the case.

To enhance immersion, the manipulations were designed to closely resemble real-world experiences. The act of inserting and removing blocks parallels the handling of a dollhouse, reinforcing tangible interaction and deepening player engagement [3]. The case serves as the designated area where players place blocks, enabling the system to detect each block's position.

Players follow the game flow illustrated in Fig. 2, using the components. The main menu allows players to choose whether to begin or exit the game. After the player chooses to start the game, a virtual environment reflecting the current block arrangement appears on the screen. If the objective—having all characters in a "satisfied" state—is achieved, players can return to the main menu. If not, they may swap the positions of two blocks and check the display again. This process repeats until the player successfully arranges all characters in the satisfied state. Through this interaction flow, players experience a heightened level of immersion.

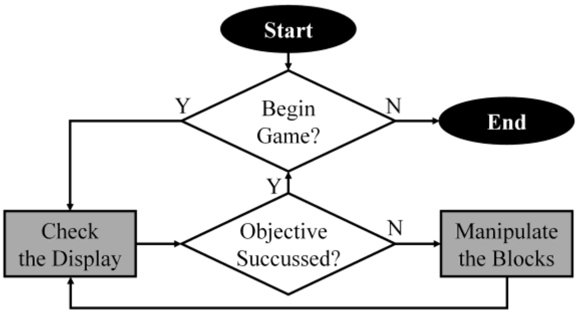


Fig. 2 Flowchart of the puzzle

III. Game Development

The development of the game system is structured into three fundamental domains: hardware, graphics, and software. These domains were executed to develop the puzzle.

A. Hardware

The six blocks and the case are powered by Arduino, an open-source physical computing platform. The external structure of the hardware was fabricated using laser-cut MDF plywood, selected for its durability and ease of sculpting. The completed blocks and case, which players use to interact with the game, are shown in Fig. 3.

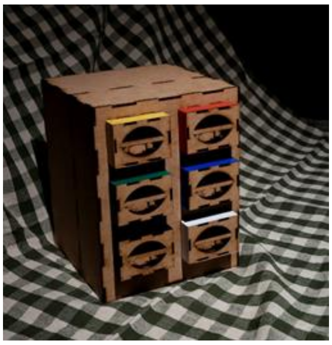
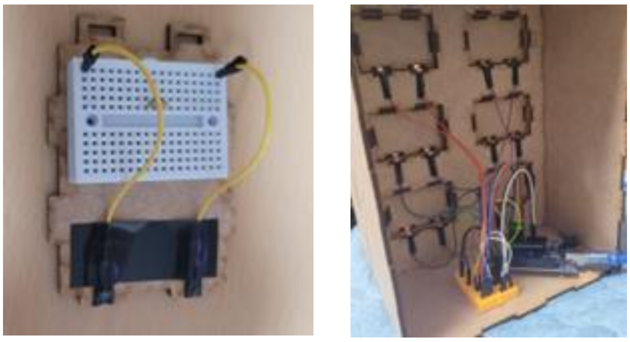


Fig. 3 Completed case containing six blocks

Each block is embedded with a resistor of a specific value, as shown in Fig. 4(a). Copper electrodes are placed on both the blocks and the case to establish an electrical circuit. Since the voltage of the circuit can be controlled by the resistor, each block's unique resistor value generates a distinct voltage signal when inserted.

This design allows the Arduino to accurately identify each block based on the voltage signal. By detecting these voltage changes, the Arduino board inside the case determines the position of each block, as shown in Fig. 4(b). Utilizing Arduino's serial communication, the software managing the game environment can receive the player's inputs.



(a) Block circuit

(b) Case circuit

Fig. 4 Arduino circuits embedded in the blocks and case

B. Graphics

All assets were created using the 3D graphics tool Autodesk Maya. To develop the puzzle, 3D models of the six characters and their house were required. As shown in Fig. 5, 3D polygons were used to construct the necessary assets.

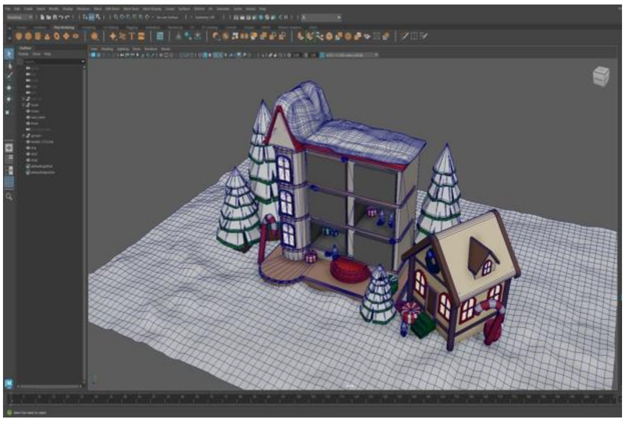


Fig. 5 Process of creating the 3D model of the house

Since the puzzle does not require high-quality, detailed graphics, we simplified the character. A single character model was created and differentiated by varying its color. Their house's wall colors were matched to the colored tape attached to the handle part of each block, allowing players to easily recognize each character and its corresponding block.

To represent the two possible states of each character, two distinct character motions were produced. Human-type joints were implanted in each character, and animations were applied accordingly, as shown in Fig. 6.

The overall aesthetic theme of the game follows a Christmas-inspired design.

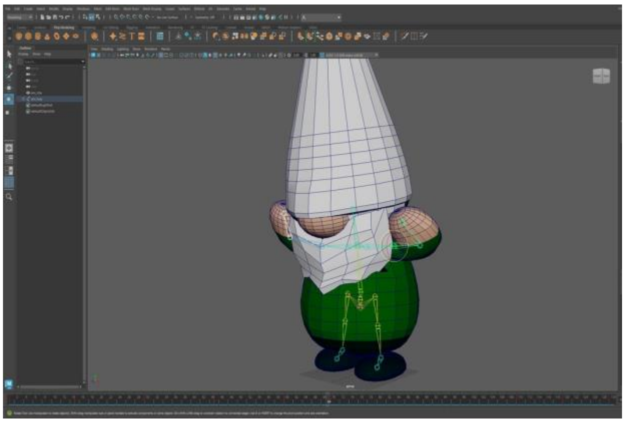


Fig. 6 Process of animating the unsatisfied character

C. *Software*

The software was developed using the Unreal Engine. Since Unreal Engine supports a visual scripting tool called Blueprint, the entire game system was implemented using this feature, as shown in Fig. 7.

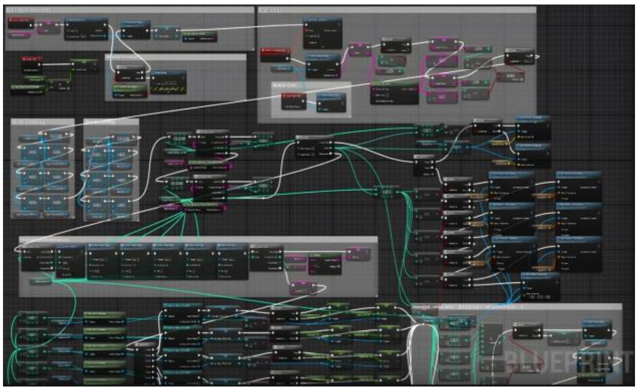


Fig. 7 Blueprint for the game system

By default, Unreal Engine does not support serial communication. Therefore, the external plugin "Serial COM" was utilized to facilitate the transmission of block position data from the Arduino board to Unreal Engine [4]. This plugin also be integrated within Blueprint.

Upon receiving data from the Arduino board via the Serial COM plugin, Blueprint pre-processes the input. The processed data is then used to assign each character to its corresponding position within the house.

Next, Blueprint calculates whether each character is in a "satisfied" or "unsatisfied" state based on a defined function. After determining each character’s state, the Blueprint triggers the appropriate animation corresponding to that state.

Finally, the Blueprint checks whether all characters are placed within the house and are in the "satisfied" state to determine if the player has fulfilled the objective of the puzzle. This entire process repeats until the player successfully finds a block arrangement where every character is in the house and in the satisfied state.

Additionally, as shown in Fig. 8, an interactive start menu and a gameplay timer were implemented. The start menu allows users to initiate or exit the game. The gameplay timer tracks and displays the session duration upon puzzle completion. The timer is shown in the top right corner of the start menu after gameplay.

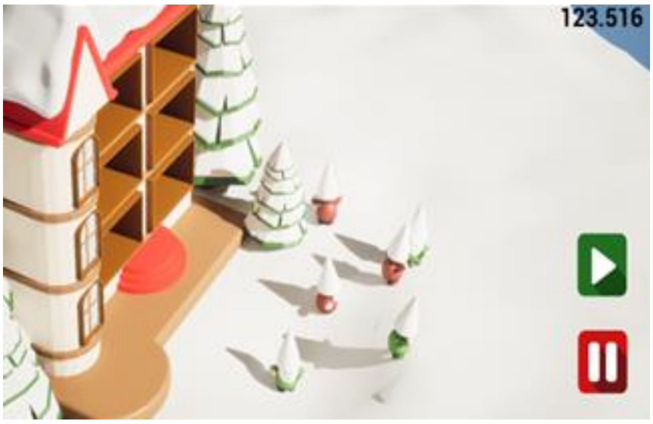


Fig. 8 Start menu with gameplay timer

IV. Experiment

The gameplay screen is presented in Figure 9. It can be observed that the arrangement of characters within the game changes according to the configuration of the block.

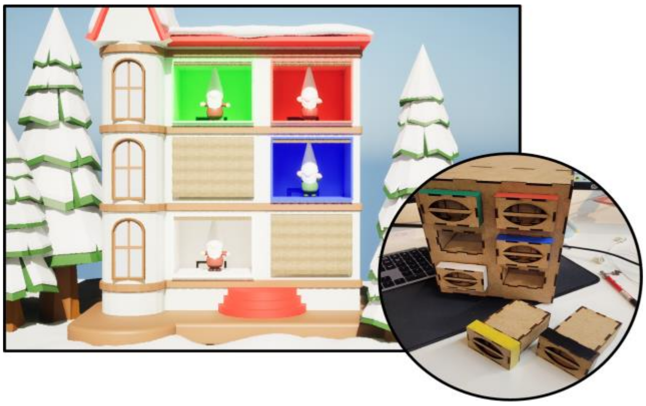


Fig. 9 Gameplay screen based on block arrangement (bottom-left circle)

To evaluate the game, the following environment was set up as shown in Fig. 10. The game was conducted on a computer equipped with an Intel® i7-6700 CPU, 16 GB of RAM, and an RTX 2060 Super graphics card. Before beginning the gameplay, participants received brief instructions on how to insert and remove the blocks.

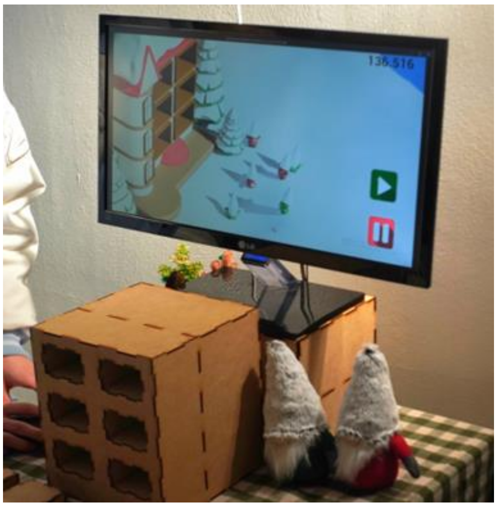


Fig. 10 Game evaluation environment

After the gameplay, brief verbal feedback was collected from the participants. Overall, they expressed satisfaction with the game manipulation system using blocks. Additionally, they noted that although the game appeared simple, its unexpected difficulty enhanced their enjoyment.

V. Conclusions

We applied tangible interaction to the puzzle game by utilizing physically manipulable blocks. Through this approach, we aimed to overcome the inherent limitations of traditional games, bridging the gap between physical and virtual worlds. Our findings suggest that tangible interaction enhances player engagement by providing a more immersive and intuitive user experience.

Future research will focus on evaluating the effectiveness of tangible interaction. Rigorously designed qualitative evaluation metrics will be used to systematically assess the effectiveness of the proposed manipulation. It will primarily examine whether users can intuitively and easily access the game, as well as whether tangible interaction improves cognitive engagement and immersion.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176, Contribution Rate: 50%) and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2023 (Project Name: Development of digital abusing detection and management technology for a safe Metaverse service, Project Number: RS-2023-00227686, Contribution Rate: 50%)

References

[1] G. Tavinor, *The Art of Videogames*, 1st ed. Chichester, UK: Wiley-Blackwell, 2009.

[2] M. Danesi, *The Puzzle Instinct: The Meaning of Puzzles in Human Life*, 1st ed. Bloomington, IN: Indiana University Press, 2002.

[3] S. D. Yang, “A study on player's immersion by difference of input control devices in computer games,” *J. Korea Game Soc.*, vol. 10, no. 1, pp. 35-46, Feb. 2010.

[4] (2024) Unreal_Engine_SerialCOM_Plugin. [Online]. Available: https://github.com/videofeedback/Unreal_Engine_SerialCOM_Plugin

Advancing Pediatric Developmental Delay Detection via Facial Data and AI with Meta Quest Pro

Ahsan Aziz[1], Yongwon. Cho[2] and Yunyoung Nam[2]
[1]Department of ICT Convergence, Soonchunhyang University, Asan 31538, Korea
[2]Department of Computer Science and Engineering, Soonchunhyang University, Asan, 31538, Korea
*Contact: ynam@sch.ac.kr

Abstract — **Developmental delay is when a child does not reach developmental milestones at the expected age in areas such as motor skills, speech, cognition, or social interaction. Some types of it are Cognitive Delay, Speech and Language Delay, Motor Delay, and Social/Emotional Delay, with common causes such as Genetic conditions, Premature birth or low birth weight, Neurological disorders, and Environmental factors. The equipment we used was Meta Quest Pro, an experimental design of an immersive VR game scene for task-based interactions and simulated cognitive and motor skill challenges. The data was collected as Facial feature data, including micro-expressions, muscle movements, and eye-tracking information, which was a Time-series data capturing dynamic changes in facial gestures. After preprocessing the following data, the primary path leverages the Minimum Redundancy Maximum Relevance (MRMR) method to extract and select the most relevant features, while the second path employs a Transformer for automated feature extraction. Both feature sets are subsequently classified using machine learning classifiers to assess their effectiveness. As by using MRMR approach we have achieved 86.4% accuracy on coarse tree with ratio of 80:20 and 95.6% achieved on quadratic svm with ratio of 60:40. Furthermore, transformer-based extraction we have achieved Quadratic SVM classifier achieved the result accuracy of 86.4%, and the Fine KNN classifier in 60:40 achieved the accuracy of 95.5%. Our study can contribute to developing affordable and accessible screening tools for children worldwide.**

I. INTRODUCTION

In an increasingly digitalized world, the fusion of technology and human emotion has emerged as a promising frontier, offering potential applications across diverse domains, including healthcare, education, entertainment, and human-computer interaction. Among the avenues of exploration, facial emotion recognition stands out as a potent tool for comprehending and enriching human experiences. The ability to decipher and respond to human facial state holds immense significance not only for creating more empathetic and responsive technology but also for its profound implications in fields such as psychology, marketing, and artificial intelligence.

Developmental delay is when a child does not reach developmental milestones at the expected age in areas such as motor skills, speech, cognition, or social interaction. Some of the types of developmental delays are stated as: Cognitive Delay – Difficulty in learning, problem-solving, and memory, Speech and Language Delay – Challenges in understanding or using spoken language, Motor Delay – Issues with fine or gross motor skills (e.g., walking, grasping objects), Social and Emotional Delay – Difficulty interacting with others, recognizing emotions, or forming relationships. The common causes for these delays are Genetic conditions (e.g., Down syndrome, Fragile X syndrome), Premature birth or low birth weight, Neurological disorders (e.g., cerebral palsy, autism spectrum disorder), and Environmental factors (e.g., malnutrition, lack of stimulation, exposure to toxins).

Conventional facial state recognition methods primarily rely on analyzing images and videos, extracting insights from two-dimensional representations of facial expressions. Nevertheless, these approaches possess inherent limitations in capturing the intricacies and subtleties of human emotions, which often manifest through the three-dimensional dynamics of facial features. This is where Virtual Reality (VR) comes into play—a technology that has revolutionized how we engage with digital content and, more notably, offers a distinctive avenue for the observation and analysis of emotions in an immersive and genuine context. The main objective of our study consists of the following points:

- Facial Data Acquisition & Feature Extraction (Capture and analyze children's facial data using advanced tracking technology to extract critical features related to developmental delays). Intelligent Classification Model Development (Design and implement a robust AI-driven classification model to accurately distinguish between typically developing children and those with developmental delays). Performance Evaluation & Validation (Assess and validate the classification framework's accuracy, reliability, and real-world applicability through rigorous testing and benchmarking). Early Screening & Data-Driven Decision Support (AI-powered diagnostic tool to assist early detection efforts, supporting clinicians and caregivers in timely intervention planning).
- VR-based setup is designed to create an immersive and controlled experimental environment. This setup facilitates real-time interaction and data acquisition while ensuring standardized conditions for all participants, it also enables the collection of dynamic behavioral responses, enhancing the robustness of the dataset.
- Collecting data of facial movement and expression using Meta Quest Pro.
- The preprocessing techniques are applied, which include handling missing values to prevent biases in the dataset, filtering out corrupted or incomplete data that may introduce inconsistencies, and normalizing and standardizing features to ensure uniform data distribution.
- On the raw data selection we will select the data which could be relevant for our classification, and on this step two parallel paths will be continuing first one will be using Minimum Redundancy Maximum Relevance (MRMR) and the other path will use the Transformer architecture for extraction of features and both paths will be classified by the machine learning classifiers at the end and then evaluated for the results.

- For the first path, the Minimum Redundancy Maximum Relevance (MRMR) algorithm optimizes feature selection, improving computational efficiency and enhancing the classification model's performance.
- Secondly, we will be using a Transformer-based feature extraction approach for the facial points dataset. This architecture consists of multiple layers that learn hierarchical feature representations from the input data.
- The selected features from both paths will be classified by the machine learning classifiers separately and evaluated at the end.

Once validated, the trained model is integrated into a real-world application for further assessment. deployment involves testing in a practical VR environment to evaluate its effectiveness under real-time conditions, and the system's adaptability to new user data is analyzed to ensure long-term usability and scalability.

II. RELATED WORK

Distress emotions in very young children are manifest in vocal, facial, and bodily cues. Moreover, children with different developmental conditions (i.e. autistic disorder, AD; developmental delay, DD; typically developing, TD) appear to manifest their distress emotions via different channels. In [2] their research has shown that children with different developmental conditions (AD, DD, TD) exhibit significant differences in emotional expression across facial, vocal, and bodily cues, and facial expressions, as a core dimension of emotional expression, can serve as an objective indicator of developmental status through feature analysis. Building on this foundation, this study integrates VR-based tasks and machine learning models to extract and classify dynamic facial features, overcoming the limitations of subjective human judgment. The study design includes 18 children (18 months old) divided into 3 groups: AD, DD, TD, with 42 female adults assessed distress and typicality. The main methodology consists of video clips of crying that were modified to isolate specific cues such as (Vocal Cues, Facial Cues, and Bodily Cues). their findings were that distress and typicality judgments varied across AD, DD, and TD groups. Cues from children with AD were judged as more atypical and distressed than DD and TD. Some overlap in adult responses to distress cues between AD and DD groups.

In [3] they investigate the correlation between empathy and facial-based emotion simulation in Virtual Reality (VR) and examine how users' ability to mimic avatar expressions relates to their empathy levels. they conducted a user study with 37 participants using the Meta Quest Pro VR headset and used the Facial Action Coding System (FACS) to capture 63 micro expressions during facial expression simulations then assessed empathy using the Interpersonal Reactivity Index (IRI) questionnaire. Authors designed a VR environment where participants simulated expressions based on avatar cues and analyzed the most recurrent micro expressions and their link to users' empathy. The dataset was collected from live user interactions in a controlled VR setting using Meta Quest Pro sensors. Their study found a statistically significant correlation between empathy and the ability to simulate emotions in VR, which provided insights into the prevalence of micro expressions across seven distinct emotions and proposed applications in mental health and emotional well-being, particularly for VR-based therapy and training. In [5] they develop a non-invasive system for predicting depression, anxiety, and stress (DASS) levels based on facial expressions and utilize machine learning to offer real-time analysis for mental health assessments. Their approach predicted DASS levels (Normal, Mild, Moderate, Severe, or Extremely Severe) based on facial expressions and evaluated the effectiveness of the system on intrasubject and intersubject testing methodologies. The Dataset they used was tested on AVEC 2014 and ANUStressDB datasets (including Cohn-Kanade (CK+), MMI, JAFFE, and other facial expression databases for AU classification). They have achieved 87.2% accuracy for depression, 77.9% for anxiety, and 90.2% for stress. And showed 93% accuracy in distinguishing healthy individuals from those with Major Depressive Disorder (MDD) or PTSD. Identified novel correlations between FACS AUs and DASS levels, increasing predictive accuracy by 5%.

III. PROPOSED METHOD

This section will discuss the proposed methodology for pediatric development delay classification using the facial points gathered from the VR device. Below is the proposed architecture shown in Figure 1.

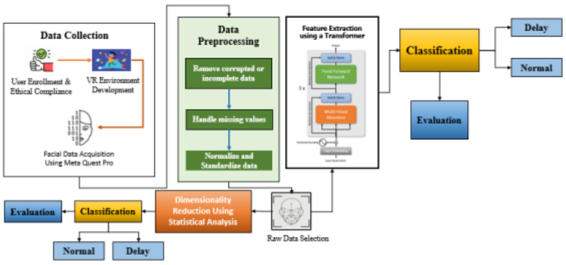

Data Collection
User Enrollment & Ethical Compliance
VR Environment Development
Facial Data Acquisition Using Meta Quest Pro
Data Preprocessing
Remove corrupted or incomplete data
Handle missing values
Normalize and Standardize data
Feature Extraction using a Transformer
Add & Norm
Feed Forward Network
3 x
Multi-Head Attention
Classification
Delay
Normal
Evaluation
Evaluation
Classification
Dimensionality Reduction Using Statistical Analysis
Raw Data Selection
Normal
Delay

Figure 1 Proposed Flow Diagram

A. *Data Collection and Preprocessing*

Initially, participants will undergo an enrollment process that ensures compliance with ethical guidelines, involves obtaining informed consent, ensuring privacy protection, and adhering to institutional and regulatory requirements. Ethical approval is obtained from relevant review boards to ensure the responsible handling of sensitive user data. Then a VR-based setup is designed to create an immersive and controlled experimental environment. This setup facilitates real-time interaction and data acquisition while ensuring standardized conditions for all participants, it also enables the collection of dynamic behavioral responses, enhancing the robustness of the dataset. Collecting data of facial movement and expression using Meta Quest Pro, a high-fidelity device equipped with advanced tracking capabilities, the system captures multimodal facial and head movement features, providing a rich dataset for subsequent analysis, and captured data include parameters such as eye movements, head orientation, and muscle activations.

Ensuring the quality and consistency of the dataset, the preprocessing techniques are applied, which include handling missing values to prevent biases in the dataset, filtering out corrupted or incomplete data that may introduce inconsistencies, and normalizing and standardizing features to ensure uniform data distribution, improving model convergence and stability during training.

B. *Feature Extraction and Selection*

On the raw data selection, we will select the data which could be relevant for our classification, and on this step two parallel paths will be continuing first one will be using MRMR and the other path will use the Transformer architecture for extraction of features and both paths will be classified by the machine learning classifiers at the end and then evaluated for the results. For first path the raw dataset that we have collected has high dimensionality, so feature selection methods are employed to retain the most informative attributes while removing redundant or irrelevant features, the Minimum Redundancy Maximum Relevance (MRMR) algorithm is used to optimize feature selection, improving computational efficiency and enhancing the performance of the classification model. Secondly, a Transformer is employed for automatic feature extraction, This architecture consists of multiple layers that learn hierarchical feature representations from the input data.

C. *Machine Learning (ML)*

With the standardized and categorized features in hand, we proceeded to employ various machine learning algorithms, with a predominant focus on Support Vector Machines (SVM). SVMs, renowned for their versatility and proficiency in classification tasks, were instrumental in the context of binary classification. These algorithms were trained on the pre-processed data, leveraging the extracted facial features to make predictions regarding the emotional states expressed by the subjects. For evaluation we will be using standard evaluation metrics, including accuracy, precision, recall, and F1-score. Cross-validation techniques are employed to ensure the robustness and generalizability of the model. Comparative analysis with baseline models is conducted to validate performance improvements.

Once validated, the trained model is integrated into a real-world application for further assessment, deployment involves testing in a practical VR environment to evaluate its effectiveness under real-time conditions, and the system's adaptability to new user data is analyzed to ensure long-term usability and scalability.

IV. RESULTS AND DISCUSSION

A. *User Enrollment & Ethical Compliance*

Before collecting children's facial data for developmental delay analysis, it is essential to ensure legal and ethical compliance. Figure 2 shows the flow chart with details, and also the steps are explained below the figure.

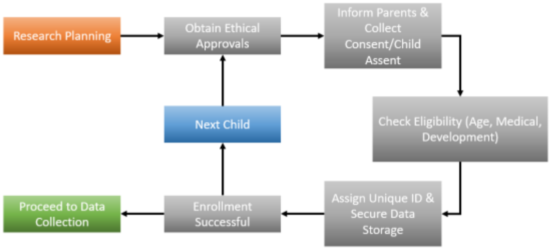

Research Planning
Obtain Ethical Approvals
Inform Parents & Collect Consent/Child Assent
Check Eligibility (Age, Medical, Development)
Assign Unique ID & Secure Data Storage
Enrollment Successful
Proceed to Data Collection
Next Child

Figure 2 Flow Chart for User Enrollment & Ethical Compliance

- Ethical Approvals

Obtain ethical clearance from relevant institutions and research boards.

- Parental Consent & Child Assent

Inform parents/guardians about the research objectives, risks, and benefits and collect signed consent forms allowing children's participation.

- Eligibility Check & Participant Screening

Verify children's age, medical history, and developmental status to ensure they meet the study criteria and conduct preliminary assessments for eligibility.

- Privacy & Anonymization

Assign unique ID codes instead of using names or identifiable data, and encrypt and securely store sensitive information.

- User Enrollment & Secure Database Management

Register participants into a secure database and generate QR codes or digital IDs for efficient tracking.

B. *Dataset*

To comprehensively understand human emotional state in immersive contexts, we leveraged VR technology as a potent tool for data collection. Our study involved child subjects actively participating in different emotional state elicitation sessions within the VR environment. The VR device captured their facial expressions and responses as they viewed and reacted to these visual stimuli. Table 1 is an explanation of all the facial points that we have used to gather the dataset.

TABLE 1 Facial Points and Their Representation

Facial Points	Action
BROW_LOWERER_L / BROW_LOWERER_R	These refer to muscles that lower the eyebrows.
CHEEK_PUFF_L / CHEEK_PUFF_R	Muscles that cause the cheeks to puff out. They are located on the sides of the face.
CHEEK_RAISER_L / CHEEK_RAISER_R	Muscles that raise the cheeks. They are located on the sides of the face.
CHEEK_SUCK_L / CHEEK_SUCK_R	Muscles that create a sucking or hollowing effect on the cheeks. They are located on the sides of the face.
CHIN_RAISER_B / CHIN_RAISER_T	Muscles that raise the chin, are typically associated with facial expressions involving the lower face.
DIMPLER_L / DIMPLER_R	These are muscles that create dimples on the cheeks, usually when someone smiles. They are located on the sides of the face.
EYES_CLOSED_L / EYES_CLOSED_R	Muscles are responsible for closing the eyes. They are located around the eyes.
EYES_LOOK_DOWN_L / EYES_LOOK_DOWN_R	Muscles that control the movement of the eyes when looking downward. They are located around the eyes.
EYES_LOOK_LEFT_L / EYES_LOOK_LEFT_R / EYES_LOOK_RIGHT_L / EYES_LOOK_RIGHT_R	Muscles are responsible for eye movement in different directions. They are located around the eyes.
EYES_LOOK_UP_L / EYES_LOOK_UP_R	Muscles that control eye movement when looking upward. They are located around the eyes.
INNER_BROW_RAISER_L / INNER_BROW_RAISER_R	Muscles that raise the inner part of the eyebrows. They are located above each eye.
JAW_DROP	Muscles that cause the jaw to drop. Located in the jaw area.
JAW_SIDEWAYS_LEFT / JAW_SIDEWAYS_RIGHT	Muscles that move the jaw sideways. Located in the jaw area.
JAW_THRUST	Muscles that thrust the jaw forward. Located in the jaw area.

LID_TIGHTENER_L / LID_TIGHTENER_R	Muscles are responsible for tightening the eyelids. They are located around the eyes.
LIP_CORNER_DEPRESSOR_L / LIP_CORNER_DEPRESSOR_R	Muscles that depress the corners of the lips are associated with expressions of sadness or disapproval. They are located around the mouth.
LIP_CORNER_PULLER_L / LIP_CORNER_PULLER_R	Muscles that pull the corners of the lips backward and upward are associated with expressions of joy or amusement. They are located around the mouth.
LIP_FUNNELER_LB / LIP_FUNNELER_LT / LIP_FUNNELER_RB / LIP_FUNNELER_RT	These refer to muscles that create funneling or puckering of the lips, typically found in the corners of the mouth.
LIP_PRESSOR_L / LIP_PRESSOR_R	Muscles that press the lips together are found around the mouth.
LIP_PUCKER_L / LIP_PUCKER_R	Muscles that create a puckering effect of the lips, located around the mouth.
LIP_STRETCHER_L / LIP_STRETCHER_R	Muscles that stretch the lips horizontally are typically found on the sides of the mouth.
LIP_SUCK_LB / LIP_SUCK_LT / LIP_SUCK_RB / LIP_SUCK_RT	These refer to muscles that create a sucking motion, typically found at various locations around the mouth.
LIP_TIGHTENER_L / LIP_TIGHTENER_R	Muscles responsible for tightening the lips are found around the mouth.
LIPS_TOWARD	Muscles that move the lips inward or toward each other are located around the mouth.
LOWER_LIP_DEPRESSOR_L / LOWER_LIP_DEPRESSOR_R	Muscles that depress the lower lip, are typically found below the lower lip.
MOUTH_LEFT / MOUTH_RIGHT	Muscles responsible for moving the mouth to the left or right, are typically found around the mouth.
NOSE_WRINKLER_L / NOSE_WRINKLER_R	Muscles that cause the nose to wrinkle, are typically located near the nose.
OUTER_BROW_RAISER_L / OUTER_BROW_RAISER_R	Muscles that raise the outer part of the eyebrows are typically found above each eye.
UPPER_LID_RAISER_L / UPPER_LID_RAISER_R	Muscles responsible for raising the upper eyelids are found around the eyes.
UPPER_LIP_RAISER_L / UPPER_LIP_RAISER_R	Muscles that raise the upper lip, are typically located above the upper lip.

Our dataset consists of 58 normal children's scenarios in data collection, and for the development delay children group, we have 54 children scenarios in data collection. The equipment we used is Meta Quest Pro with an experimental design as an Immersive VR game scene designed for task-based interactions and simulated cognitive and motor skill challenges. The VR game task design consists of object recognition (distinguishing between various objects in a virtual space) and object manipulation (Performing precise movements to interact with objects and testing motor skills). The purpose of the task design was to capture a diverse range of facial expressions and movements, extract key cognitive and motor skill indicators, and facilitate feature extraction for the classification of normal vs. developmentally delayed children.

C. *Experimental Results*

This study evaluated the performance of different feature selection methods and classifiers on a VR children's facial points dataset. The evaluation was performed on the test set, concentrating on metrics such as accuracy, precision rate, specificity rate, sensitivity rate, F1 score and AUC. In Table 2, the results of the machine learning classifiers using the statistical technique MRMR with the train and test ratio of 80:20 and 60:40. It selects a set of features that are highly relevant to the target class and distinct from each other, making the model more efficient and potentially improving performance.

TABLE 2 Results using MRMR with ML

ML Classifiers	Accuracy	Recall Rate	Specificity Rate	Precision Rate	F1 Score	AUC
Ratio 80:20						
Cubic SVM	82.6%	1.00	0.7333	0.6667	0.800000	0.75758
Quadratic SVM	73.9%	0.8000	0.6923	0.6667	0.727273	0.77273
Fine KNN	78.3%	0.8182	0.7500	0.7500	0.782609	0.78409
Coarse Tree	**86.4%**	**0.8462**	**0.8889**	**0.9167**	**0.880000**	**0.85833**
Ratio 60:40						
Cubic SVM	91.1%	1.00	0.8462	0.8261	0.904762	0.97628
Quadratic SVM	**95.6%**	**1.00**	**0.9167**	**0.9130**	**0.954545**	**0.99012**
Fine KNN	84.1%	0.8636	0.8182	0.8261	0.844444	0.84161
Coarse Tree	93.3%	0.9545	0.9130	0.9130	0.933300	0.92885

Figure 4 shows the confusion matrix of Cubic SVM and Quadratic SVM classifier and Figure 5 shows the AUC curve of the following classifier.

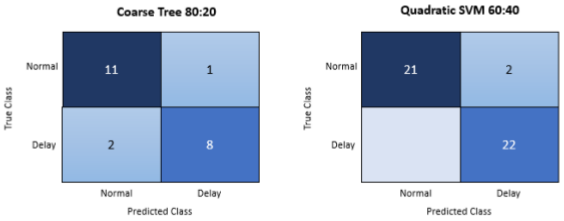


Figure 3 Confusion matrix of Coarse Tree and Quadratic SVM classifier with ratio (80:20 and 60:40)

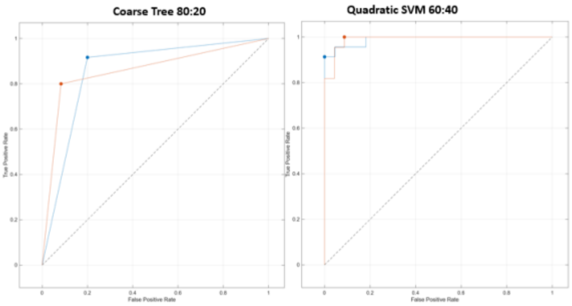


Figure 4 AUC Curve of Coarse Tree and Quadratic SVM classifier with ratio (80:20 and 60:40)

In Table 3, the results of the machine learning classifiers using a transformer with a train and test ratio of 80:20 and

60:20.

TABLE 3 Results using Transformer with ML

ML Classifiers	Accuracy	Recall Rate	Specificity Rate	Precision Rate	F1 Score	AUC
80:20						
Cubic SVM	77.3%	0.8182	0.7273	0.7500	0.782609	0.7333
Quadratic SVM	**86.4%**	**1.00**	**0.7857**	**0.7273**	**0.842105**	**0.84167**
Fine KNN	81.8%	0.9000	0.7500	0.7500	0.818182	0.8250
Coarse Tree	68.2%	0.7273	0.6364	0.6667	0.695652	0.6250
60:40						
Cubic SVM	88.6%	0.8750	0.9000	0.9130	0.893617	0.9441
Quadratic SVM	88.6%	0.9091	0.8636	0.8696	0.88889	0.89855
Fine KNN	**95.5%**	**1.00**	**0.9130**	**0.9130**	**0.954595**	**0.95652**
Coarse Tree	75.0%	0.8000	0.7083	0.6957	0.744186	0.75673

Figure 5 shows the confusion matrix, and Figure 6 shows the AUC Curve of the Quadratic SVM and Fine KNN classifier. As the following Quadratic SVM classifier achieved the result accuracy of 86.4%, and the Fine KNN classifier in 60:40 achieved the accuracy of 95.5% in this case.

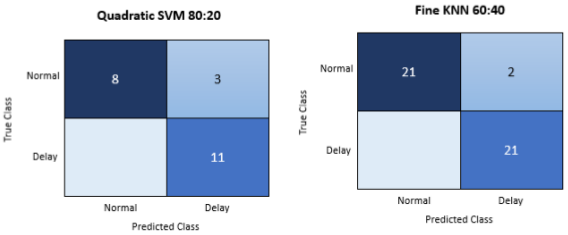


Figure 5 Confusion Matrix of Quadratic SVM and Fine KNN classifier with ratio (80:20 and 60:40)

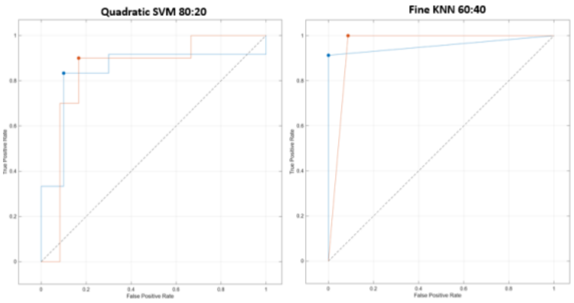


Figure 6 AUC Curve of Quadratic SVM and Fine KNN classifier with ratio (80:20 and 60:40)

V. CONCLUSION

In this study, we anticipated a dual-path approach for classifying facial movement data acquired employing Meta Quest Pro. The primary path leverages the Minimum Redundancy Maximum Relevance (MRMR) method to extract and select the most relevant features, while the second path employs a Transformer for automated feature extraction. Both feature sets are subsequently classified using machine learning classifiers to assess their effectiveness. As by using MRMR approach we have achieved 86.4% accuracy on coarse tree with ratio of 80:20 and 95.6% achieved on quadratic svm with ratio of 60:40. Furthermore, transformer-based extraction we have achieved Quadratic SVM classifier achieved the result accuracy of 86.4%, and the Fine KNN classifier in 60:40 achieved the accuracy of 95.5%.

Through rigorous preprocessing, normalization, and feature selection, the proposed framework certifies optimal data representation, enhancing classification performance. The comparative analysis between MRMR-based and Transformer-based feature extraction delivers insights into their corresponding contributions to facial movement classification. The results demonstrate the feasibility of using advanced feature extraction techniques for accurate classification in real-world applications.

Traditional diagnosis is often subjective and requires expert evaluation. AI-driven facial analysis provides a non-invasive, fast, and scalable method for early detection, our study contributes to developing affordable and accessible screening tools for children worldwide. Future work will focus on optimizing feature selection strategies, incorporating additional deep learning architectures, and expanding the dataset for improved generalization. The proposed methodology paves the way for further advancements in facial movement analysis, contributing to applications in healthcare, behavioral studies, and virtual reality interactions.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176).

VI. REFERENCE

[1] Geraets, C. N. W., S. Klein Tuente, B. P. Lestestuiver, M. Van Beilen, S. A. Nijman, J. B. C. Marsman, and W. Veling. "Virtual reality facial emotion recognition in social environments: An eye-tracking study." *Internet interventions* 25 (2021): 100432.

[2] Assessment of distress in young children: A comparison of autistic disorder, developmental delay, and typical development. http://dx.doi.org/10.1016/j.rasd.2011.02.013

[3] A user study on the relationship between empathy and facial-based emotion simulation in Virtual Reality. ACMISBN979-8-4007-1764-2/24/06 https://doi.org/10.1145/3656650.3656691

[4] Banerjee, Agnik, Onur Cezmi Mutlu, Aaron Kline, Saimourya Surabhi, Peter Washington, and Dennis Paul Wall. "Training and profiling a pediatric facial expression classifier for children on mobile devices: machine learning study." *JMIR formative research* 7 (2023): e39917.

[5] Predicting Depression, Anxiety, and Stress Levels from Videos Using the Facial Action Coding System.

[6] Recognizing Action Units for Facial Expression Analysis.

[7] Deep Facial Expression Recognition: A Survey.

[8] Virtual reality for healthcare: A scoping review of commercially available applications for head mounted displays.

[9] Bieberich, A., & Morgan,S.B.(1998). Briefreport:Affective expression in childrenwith autismorDown’ssyndrome.JournalofAutism andDevelopmental Disorders, 28, 333–338.

[10] Anonymous. 2024. Raw Data of Facial Micro-Expression Intensity in Acted Facial Expressions. https://figshare.com/s/c1536f4e9b92f0137729.

[11] CDanielBatson.2014. Thealtruismquestion: Toward a social-psychological answer. Psychology Press.

[12] Xie, Hong-Xia, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. "An overview of facial micro-expression analysis: Data, methodology and challenge." *IEEE Transactions on Affective Computing* 14, no. 3 (2022): 1857-1875.

[13] Zhao, Guoying, Xiaobai Li, Yante Li, and Matti Pietikäinen. "Facial micro-expressions: An overview." *Proceedings of the IEEE* 111, no. 10 (2023): 1215-1235.

Analyzing the Effect of Inter-Eye Distance on Eye-Related Features Extraction in Virtual Reality: A Meta Quest Pro-Based Study

Syed Ali Naqi Raza[1], Yongwon. Cho[2], Yunyoung Nam[3]
[1]*Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Republic of Korea*
[2]*Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea*
[3]*Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea*

*Contact: ynam@sch.ac.kr

Abstract— **Meta Quest Pro based eye and facial tracking technology plays an important role in the study of visual attention, cognitive processes, user interactions and human behavior analysis. However, differences in eyes anatomy, such as Inter Canthal Distance (ICD), Interpupillary Distance (IPD), and Outer Canthal Distance (OCD), may impact the accuracy & reliability of eye movement data. This study aims to evaluate the impact of inter eye distance variations on eye-movement based feature extraction in VR environments using Head Mounted Device Meta Quest Pro. The data collection process involved eye measurement (Interpupillary Distance, Inter Canthal Distance, and Outer Canthal Distance), an eye calibration test before experiment, and three experimental tasks: Fixed Gaze Task, Regular Eye Movement Task, and Irregular Eye Movement Task. The Meta Quest Pro recorded participants' eye movement while they performed VR-based tasks, followed by extraction of eye related features from the device. The dataset underwent normality tests (Shapiro-Wilk, D'Agostino-Pearson) and statistical analysis (Spearman Correlation, Kruskal-Wallis, and Mann-Whitney U Statistic) to assess differences in eye-movement correlations across Wide, Narrow, and Average eye distance groups. The results of statistical tests didn't indicate significant influence of inter eye distance on eye movement tracking accuracy. This study is important for research including learning disabilities research, child and adult behavior analysis, and human computer interaction, ensuring eye data remains accurate and reliable across different users.**

I. Introduction & Related Work

Virtual Reality (VR) has changed human-computer interaction by providing interactive and immersive experiences. It enhances user engagement in various fields such as healthcare, training simulations, gaming and research [1]. Current advancements have integrated eye movement tracking technology into HMD systems, gaining importance for applications such as gaze-based interactions, accessibility improvements, and usability enhancements [2]. By tracking user's eye movements, HMD systems can improve interaction methods, visual rendering, and allow for more intuitive user experiences[3]. By creating more realistic virtual avatars, eye movement tracking can also be utilized to enhance social interaction and teamwork in virtual reality settings [4]. Eye movement activity is a type of nonverbal communication that offers social indications that are crucial for successful in-person interactions [5]. The perceived quality of communication between human parties in virtual reality can be enhanced by using eye movement data to control the gaze behavior of a user's virtual avatar [6]. Various factors influence the quality of HMD's eye and facial tracking including the participants, deployed platform, hardware manufacturer and the experimental environment [7,8]. A preliminary assessment of the VR's eye tracking signal quality was provided by earlier research [9], which measured the spatial accuracy and precision of the device across 12 individuals in both confined and unconstrained environments. Research on eye-tracking accuracy in HMDs is still limited. Sipatchin et al. carried out a thorough analysis of the Vive Pro Eye, discussing the merits and restrictions of its eye tracker[10].

Meta Quest Pro a VR device is released in October 2022, having an inbuilt integrated eye and facial features tracker. It estimates the direction of the user's eye gaze using an infrared camera and processes gaze data in real time. The Meta Quest Pro allows eye data collection in both PC and standalone modes, facilitating versatile research applications[11].

As VR technology continues to improve, understanding the impact of differences in eye's anatomy, such as inter-eye distances, on eye-tracking accuracy is crucial[12]. Inter-eye distances, including inter canthal distance (ICD), outer canthal distance (OCD), and interpupillary distance (IPD) vary significantly among individuals and may influence the accuracy & reliability of eye-movement data extraction from HMDs. Understanding the inter eye distance impact on eye movement feature extraction is important for improving calibration and ensures stable data collection for industry applications and the research purpose. Furthermore, variations in eye anatomy can impact the accuracy of extracted eye-movement data or sometimes get the missing data. Thus, Investigating the influence of these anatomical differences on feature extraction performance is crucial for enhancing data accuracy and reliability. This study analyzes the inter eye's distance (IPD, OCD, ICD) and their impact on eye movement data extraction from Meta Quest Pro. IPD, ICD, and OCD were used because these anatomical measures directly impact the alignment between the user's eyes and the eye-tracking sensors in head-mounted displays (HMDs), potentially influencing eye movement accuracy. The data collection involve measuring eye distances (Inter Canthal Distance, Interpupillary Distance, and Outer Canthal Distance) followed by an eye callibration test before experiment starts, the experimental tasks include: Fixed Gaze Task, Regular Eye Movement Task, and Irregular Eye Movement Task. The Meta Quest Pro tracked the eye movements of participants while they engaged in the VR based tasks, followed by eye related features extraction from the Meta Quest Pro. The collected features underwent normality tests (Shapiro-Wilk, D'Agostino-Pearson) and statistical analyses (Spearman Correlation, Kruskal-Wallis, and Mann-Whitney U Statistic) to evaluate differences in eye-movement relationships among Wide, Narrow, and Average eye distance groups. The study is important for learning disabilities research, child and adult behavior analysis, and human computer interaction, ensuring eye data remains accurate and reliable across different users. To address the mentioned concerns, this study highlights following research questions:

RQ1: Does inter-eye distance significantly influence the accuracy of eye movement feature extraction in Meta Quest Pro?
RQ2: Do relationships between eye movement features (e.g., EyesLookDownL, EyesLookDownR) extracted through Meta Quest Pro remain stable across different inter-eye distances?

II. Methodology

A. Participants Selection:

We recruited 10 male participants from the SoonChunHyang University (aged between 26-30 years having no serious illness) to take part in the study. All participants had normal vision. 4 participants wore glasses during the experiement. All participants were thoroughly briefed about the tasks prior to the experiment. Furthermore, participants were asked to perform an eye calibration test to enhance the participant's experience in the VR environment.

B. Apparatus and Software:

The software used for VR Applications was Unity 2021.3.45f1. The stimulus were developed and runned on Windows 10 Operating System computer(Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz 3.70 GHz) Same computer was used to display VR content on Meta Quest Pro. The Meta Quest Pro was connected through Air Link with the PC. The data was recorded using Meta Quest Pro's builtin tracking features. Meta XR All In One SDK (UPM) Version 72.0 was used for the tracking features [13]. The Meta Quest Pro was selected for this study due to its advanced eye and facial tracking capabilities, adjustable IPD settings, and increasing use in behavioral and cognitive VR research, making it an ideal device to assess whether eye movement data remains consistent across varying inter-eye anatomical differences

C. VR Tasks:

Fixed Gaze Task:

In fixation gazed task participants were instructed to fixate on a virtual sphere which is displayed at a fixed position for 30 seconds in a virtual environment. During this period, participants were asked to blink their eyes for 20 times enabling the system to capture the Eye Closeness feature (EyesClosedL, EyesClosedR) . The virtual sphere was placed at a distance of 1 meter to maintain uniform experimental conditions. The eye openness and closeness were recorded continiously during the task. The study employed within-in subject design, in which each participant completed the same blink task. The VR camera was placed at origin (0,0,0) in the Unity's world coordinate system.

Irregular Movement Task

After the Gaze Fixation Task, participants were asked to perform Irregular Movement Task, designed for vertical and horizontal eye movements evaluation. In the task, 4 spheres were displayed for seconds sequentially in up, down, left and right direction relative to the central gaze position. The left and right spheres were positioned ±25° horizontally, the up and down spheres were placed at ±25° vertically. The participants were asked to keep their head steady and to move only the eyes. This task examined both eyes' movement patterns. Table 1 lists the extracted eye movement features.

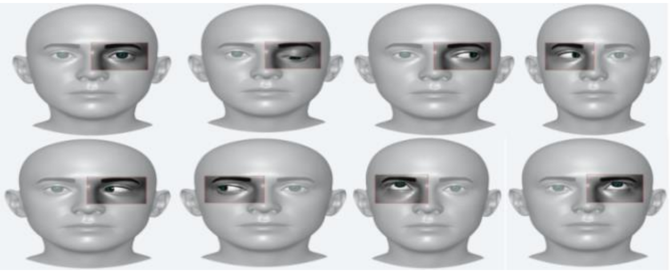


Figure 1 Sample of Extracted Eye Movement Features

TABLE I
EXTRACTED EYE MOVEMENT FEATURES

Features	Description
EyesLookLeftL	Left eye looking at left direction
EyesLookLeftR	Right eye looking at left direction.
EyesLookRightL	Left eye looking at right direction.
EyesLookRightR	Right eye looking at right direction.
EyesLookUpL	Left eye looking upwards
EyesLookUpR	Right eye looking upwards.
EyesLookDownL	Left eye looking downwards.
EyesLookDownR	Right eye looking downwards.
EyesClosedL	Closeness of left eye
EyesClosedR	Closeness of right eye

Regular Movement Task

In regular movement task participants were asked to follow a moving object as it moves in a square or rectangular path within the virtual environment. Each object was displayed for a duration of 10 secs. Similar to Fixed Gaze Task and Irregular Movement Task, this task also employed within-in subject design, in which each participant completed the same movement task. The fixed gaze task, irrergular movement task, and the regular movement task had dark background in a non reflective environment.

D. Procedure:

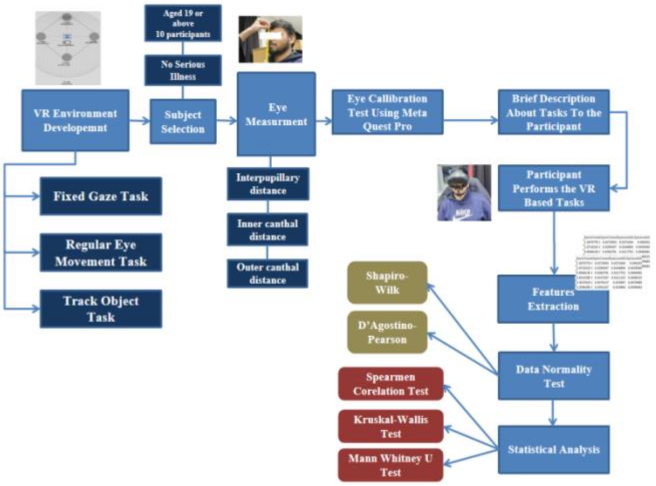


Figure 2 Overall Methodology of Study

The participants were invited one at a time to the lab for a single time visit to perform the VR task. Upon arrival, the participant's inter eye distances are calculated using a measuring tape. These measurements include Interpupillary distance, Inter Canthal Distance and Outer Canthal Distance. Table-2 defines the mentioned distances in detail.

TABLE II
DESCRIPTION OF DISTANCES BETWEEN EYES

Interpupillary Distance	The distance between the centers of the pupils of both eyes.
Inter Canthal Distance	The distance between the inner corners (medial canthi) of both eyes.
Outer Canthal Distance	The distance between the outer corners (lateral canthi) of both eyes.

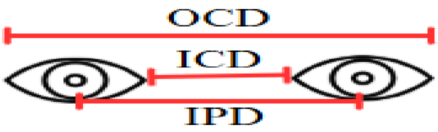


Figure 3 Description Of Distances Between Eyes

The participants' eyes distance are divided into 3 categories wide, average and narrow based on IPD, ICD and OCD distances using Z-Score.

$$Z = \frac{X - \mu}{\sigma}$$

X=individual measurement, μ=Mean of data, σ= Standard Deviation of data

After recording the eye distances participants were fitted with Meta Quest pro headset and were asked to perform eye calibration test. A detailed description of tasks was provided for the smooth conduct of experiments. During experiment the participants were asked to perform tasks discussed in section 2.3 (Fixed Gaze Task, Irregular Movement Task, Regular Movement Task). Gaze task extracted the eye closeness features, followed by Irregular and Regular Movement Tasks to extract the eye movement features (left, right,up,down gaze movement). The eye movement and eye blink data extracted from Meta Quest pro was stored in CSV files in the computer. A detailed statistical analysis was conducted to analyze the impact of inter eye distances on eye movement feature extraction. The statistical methods include normality tests, correlation analysis and non parametric tests to identify the relationships and differences between eye movement features for different eye distances groups (Wide, Average, Narrow) IPD, ICD and OCD.

The normality of extracted eye features was checked using Shapiro-Wilk test and D'Agostino-Pearson Test. Tests were conducted to determine the normal distribution of eye movement features. The Spearman Corelation Test measured the relationships between eye movement features particularly evaluating the consistency of eye closeness of both eyes, horizontal eye movement correlations and vertical gaze correlation. Kruskal Wallis Test was performed to compare and to determine the significant differences between eye movement features correlation across different eye distance groups (IPD, ICD and OCD). To further investigate the pairwise differences between across eye distance groups, Mann-Whitney U test was performed. Figure-2 shows the overall flow of study.

III. RESULTS

A. Spearman Correlation Analysis:

The Spearman Correlation Analysis was conducted to analyze the relationship between extracted eye movement features across Inter Pupillary Distance (IPD), Inter Canthal Distance (ICD), and Outer Canthal Distance (OCD).

Interpuiplary Distance:

In Inter-Pupillary Distance, synchronization of downward gaze was highly consistent (0.95) in all IPD groups (Wide, Narrow and Average) showing that downward gaze tracking was not affected by pupil distance. Similarly, upward and downward gaze movements were negatively correlated in narrow, wide and average IPD groups (-0.72) indicating the natural opposition in vertical eye movement, regardless of inter papillary distance. Left and Right gaze correlations were highest in the Wide IPD group (0.85), individual with larger

IPD exhibited more balanced gaze movements. The narrow IPD distance showed slightly lower left-right gaze coordination (0.80). Blink Synchronization was high in the IPD group (0.51), followed by Narrow group (0.48). This showed that individuals with larger inter pupillary distance have more synchronized blinking behavior but there was no significant difference with narrow and average grouped IPD. The result indicates that variations in Inter Pupilary distances have minimal impact on accuracy of eye movement feature extraction in meta Quest Pro. Figures 4,5,6 shows the results of Spearmen Correlation for average, narrow and wide Interpupilary distances.

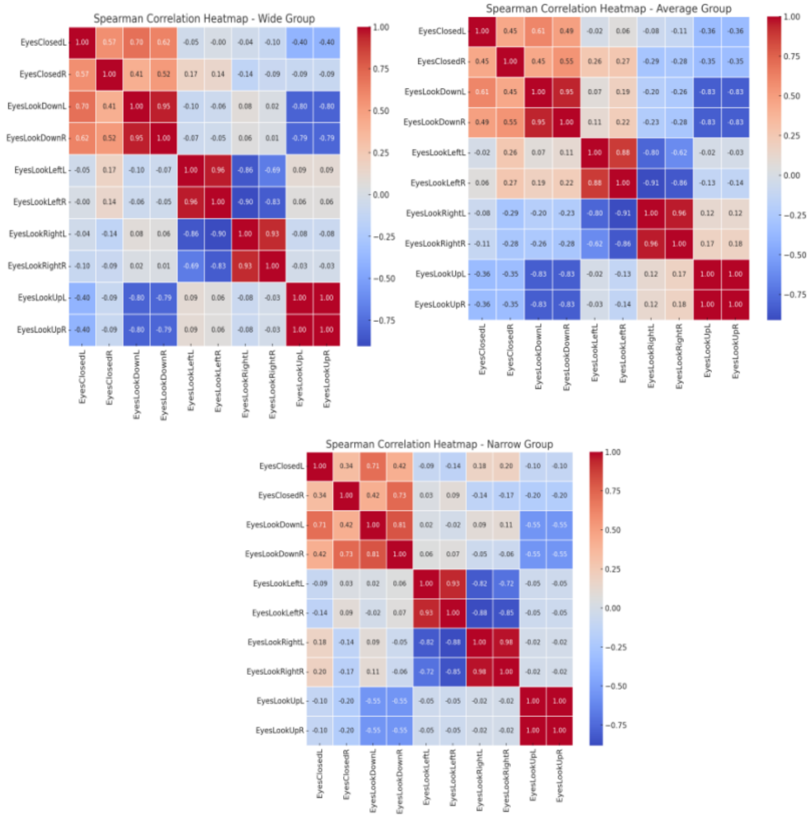


Fig. 4,5,6 Spearman Correlation for IPD Wide, Average and Narrow groups.

Inter-Canthal Distance:

The Inter-Canthal Distance is the horizontal distance between the inner corner of both eyes. The analysis of eye movement features extracted from Meta Quest Pro in relation to ICD revealed that the synchronization of EyesLookDownL and EyesLookDownR features remained highly stable (0.95) across all ICD groups, indicating that the variations in the inner eye spacing do not affect the downward gaze movement in Meta Quest Pro. The correlation between left right gaze (EyesLookLeftL vs. EyesLookLeftR) was highest in the Wide ICD group (0.96), indicating highly stable synchronization. The Narrow ICD group showed slightly lower correlations (0.93). Furthermore, upward and downward eye movements showed negative correlations in all groups (Wide, Average and Narrow) that indicates natural opposition in vertical eye movement. Blink synchronization difference was not statistically significant, the wide ICD showed more stability (0.50) as compared to narrow and average (0.49) ICD distance.

Overall, the Spearman Correlation analysis for variations in Inter-Canthal Distance suggests that ICD variations do not significantly affect eye movement tracking or the accuracy of eye based features extraction from Meta Quest Pro. Figures 7,8,9 shows the spearman correlation of wide, average and narrow Inter Canthal Distances.

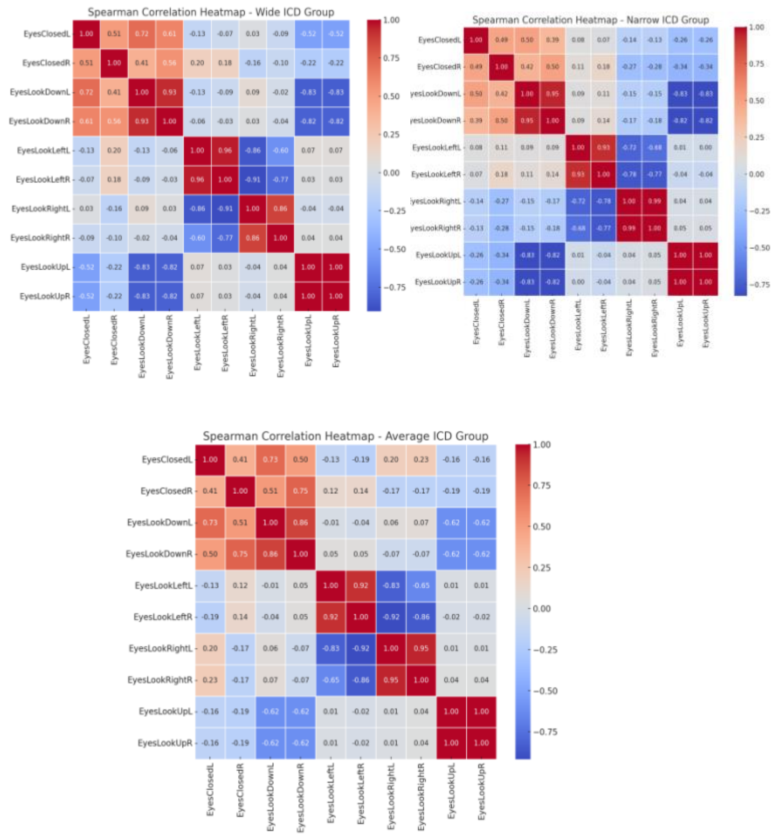


Fig. 7,8,9 Spearman Correlation for ICD Wide, Average and Narrow groups

Outer-Canthal Distance

Outer Canthal Distance (OCD) is referred as the distance between outermost corners of eyes. It is one of the important parameter when extracting eye movement features from Meta Quest Pro. The analysis of eye movement features across OCD groups revealed that downward gaze tracking is not influenced by outer canthal distance as downward gaze synchronization for both eyes (EyesLOOkDownL and EyesLookDownR) showed consistent correlation (0.95) across all OCD groups. Similarly, upward and downward eye movements exhibits negative correlations (-0.70 to -0.72) across all the groups of OCD. Thus aligning with the natural eye movement behaviour. Furthermore, Blink coordination (EyesClosedL and EyesClosedR) shown to be highest in average OCD group (0.58), on the other hand narrow group showed more variability (0.49) as compared to average and wide groups. The wide outer canthal distances showed most balanced horizontal eye movement (0.92 for both left and right movement), the average and narrow OCD participants showed slightly lower stability (0.90 & 0.83). These results show that variations in OCD distances do not significantly influence the eye movement tracking performance, however, minor variations in horizontal gaze movement and blink synchronization were appear. Figures 10,11,12 shows the spearman correlation of wide, average and narrow Outer Canthal Distances.

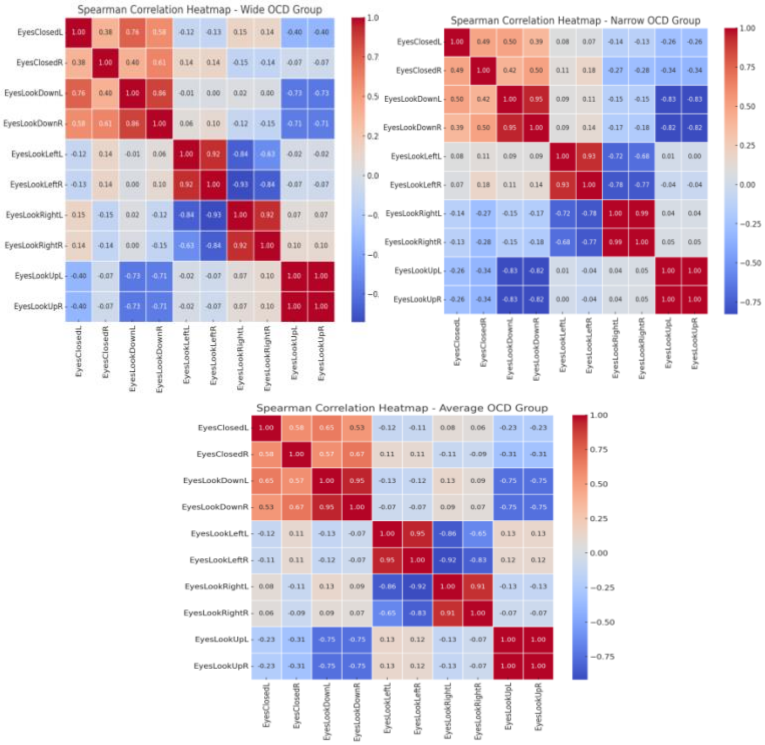


Fig. 10,11,12 Spearman Correlation for OCD Wide, Average and Narrow groups

B. Kruskal Wallis Test & Mann Whitney U Test:

To further investigate the influence of variations in inter-pupillary distance, inter-canthal distance, and outer-canthal distance on the extraction of eye movement features from Meta Quest Pro and to analyze about missing or incorrect data Kruskal Wallis test was conducted. Kruskal Wallis test checked the statistically significant differences in eye movement feature correlations between Wide, Narrow and Average groups of IPD, ICD and OCD. The results showed that p value for IPD was 0.92, ICD had 0.910, and OCD 0.861. Since all the p values were greater than 0.05, the results indicate that there were no statistically significant differences in eye movement feature correlations between Wide, Narrow and Average groups. Table-3 summarizes the results for Kruskal Wallis Test. These findings show that variations in inter eye distance do not significantly influence eyes movement feature extraction in Meta Quest Pro.

TABLE III

KRUSKAL WALLIS TEST RESULTS

Group	Statistic	P-value
Inter Pupillary Distance	0.165	0.920
Inter Canthal Distance	0.176	0.91
Outer Canthal Distance	0.298	0.861

For further validation, Mann-Whitney U test was conducted to compare the correlation distributions between all pair of groups for IPD, OCD and ICD. The groups include Wide vs. Narrow, Wide vs. Average, narrow vs. Average. All comparisons between groups had p-value more than 0.05. This indicates no significant differences between the groups. Thus, Mann-Whitney U test supports the Kruskal Wallis results, by confirming that inter eye distances do not significantly influence the eye movement feature extraction from Meta Quest Pro as the distributions do not vary significantly across Wide, Narrow and Average groups. Table 4 summarizes the Mann Whitney U test results.

TABLE IV

MANN-WHITNEY U TEST RESULTS

Groups	Inter Pupillary Distance	Inter Canthal Distance	Outer Canthal Distance
Wide vs. Narrow (U)	994.0	1011.0	1062.0
Wide vs Narrow (P value)	0.884	0.993	0.692
Wide vs Average (U)	1041.0	968.0	1043.0
Wide vs Average (P value)	0.821	0.722	0.808
Narrow vs Average (U)	1064.0	967.0	948.0
Narrow vs Average (p-value)	0.680	0.716	0.605

IV. CONCLUSIONS

This study analyzed the impact of variations in inter-eye distance (inter-pupillary distance, inter-canthal distance, and outer canthal distance) on accuracy of eye movement features extraction (EyesLookLeftL, EyesLookLeftR, EyesLookRightL, EyesLookRightR, EyesLookUpL, EyesLookUpR, EyesLookDownL, EyesLookDownR) using Meta Quest Pro. Fixed gaze task, regular movement task and irregular movement task were performed by participants in VR based environment to evaluate the consistency of eye movement across different eye anatomical structures. The study answered following research questions:

RQ1: Does inter-eye distance significantly impact the accuracy of eye movement feature extraction in Meta Quest Pro?

RQ2: Do relationships between eye movement features (e.g., EyesLookDownL, EyesLookDownR) extracted through Meta Quest Pro remain stable across different inter-eye distances?

To answer the mentioned research questions, Spearman correlation analysis, Kruskal-Wallis test, and Mann-Whitney

U tests were conducted. These statistical analysis tests compared eye movement features across wide, narrow and average groups of IPD, ICD and OCD distances. The results didn’t show significant influence of inter eye distance on eye movement tracking accuracy.These findings demonstrate that eye-movement relationship remain stable across different eye distance (IPD, ICD, OCD) groups (wide, average, narrow) , means eye movement features extracted from Meta Quest Pro maintains consistent accuracy regardless of differences in eyes distance. This study is important for human computer interaction analysis, child and adult behavior analysis, and learning disabilities research ensuring eye data remains accurate and reliable across different users. In our study we used the Meta Quest Pro, in future the study can be extended by using different VR headsets for more generalized analysis. The study involved limited participants that allowed detailed observation, high level of experimental precision and minimized variability. The future research can include diverse participants like children.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176)

References

[1] B. J. Hou, Y. Abdrabou, F. Weidner, and H. Gellersen. 2024. Unveiling Variations: A Comparative Study of VR Headsets Regarding Eye Tracking Volume, Gaze Accuracy, and Precision. 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 650-655.

[2] Z. Huang, G. Zhu, X. Duan, R. Wang, Y. Li, S. Zhang, and Z. Wang. 2024. Measuring eye-tracking accuracy and its impact on usability in Apple Vision Pro. arXiv preprint arXiv:2406.00255.

[3] Tobii. 2022. Eye tracking in XR — 2022 wrap-up and what's ahead. Tobii Blog. Retrieved from https://www.tobii.com/blog/eye-tracking-in-xr-the-2022-wrap-up-and-innovations-on-the-horizon.

[4] A. K. Mutasim, A. U. Batmaz, and W. Stuerzlinger. 2021. Pinch, click, or dwell: Comparing different selection techniques for eye-gaze-based pointing in virtual reality. Proceedings of the ACM Symposium on Eye Tracking Research & Applications (ETRA) Short Papers, 1–4.

[5] A. S. Fernandes, T. S. Murdison, and M. J. Proulx. 2023. Leveling the playing field: A comparative reevaluation of unmodified eye tracking as an input and interaction modality for VR. IEEE Transactions on Visualization and Computer Graphics, 29(5), 2269–2279.

[6] A. S. Fernandes, T. S. Murdison, and M. J. Proulx. 2023. Leveling the playing field: A comparative reevaluation of unmodified eye tracking as an input and interaction modality for VR. IEEE Transactions on Visualization and Computer Graphics, 29(5), 2269–2279.

[7] BK. Holmqvist, M. Nyström, and F. Mulvey. 2012. Eye tracker data quality: What it is and how to measure it. Proceedings of the Symposium on Eye Tracking Research & Applications, 45–52.

[8] A. J. Hornof and T. Halverson. 2022. Cleaning up systematic error in eye-tracking data by using required fixation locations. Behavior Research Methods, Instruments, & Computers, 34(4), 592–604.

[9] S. Wei, D. Bloemers, and A. Rovira. 2023. A preliminary study of the eye tracker in the Meta Quest Pro. Proceedings of the ACM International Conference on Interactive Media Experiences (IMX '23), 216-221.

[10] A. Sipatchin, S. Wahl, and K. Rifai. 2020. Accuracy and precision of the HTC Vive Pro eye tracking in head-restrained and head-free conditions. Investigative Ophthalmology & Visual Science, 61(7), 5071–5071.

[11] Meta. Move Eye Tracking in Unity. Meta Developers. Accessed: Mar. 4, 2025. [Online]. Available: https://developers.meta.com/horizon/documentation/unity/move-eye-tracking/.

[12] S. Wei, D. Bloemers, and A. Rovira. 2023. A preliminary study of the eye tracker in the Meta Quest Pro. Proceedings of the ACM International Conference on Interactive Media Experiences (IMX '23), 216-221.

[13] Meta. Meta XR SDK - All-in-One. [Online]. Available: https://developers.meta.com/horizon/downloads/package/meta-xr-sdk-all-in-one-upm/. Accessed: Mar. 6, 2025.

Teleoperated Haptic Robot Arm Using ESP32

L. Saing[1], N. Khiev[1], K.H. Kim[1,*]
Department of Telecommunication and Electronic Engineering, Royal University of Phnom Penh, Phnom Penh, Cambodia
*Contact: flyworld7@gmail.com

Abstract— **In this paper, we describe the development of a system capable of mimicking human motion and performing various tasks using the ESP32 microcontroller. The system consists of two different embedded systems: one on the controller and another on the robot arm. Both systems use a real-time operating system (RTOS) to manage multiple tasks and resources with priority levels. The system on the controller consists of three inertial measurement units (IMUs) mounted on different parts of the user's arm, allowing the user to control the robot wirelessly using Bluetooth communication. The system on the robot arm uses a motion planning mechanism to map the received human motion data and use it to control the robot arm. It uses Spherical Linear Interpolation (SLERP) to ensure precise joint control with smooth orientation transitions. To avoid unexpected behavior and erratic motion, the sensors on the controller need to be calibrated in a specific posture. This system also includes a closed-loop mechanism that provides users with physical feedback from the robots to the controller. This approach has been demonstrated using a seven degrees of freedom (DOF) robot arm, which showcases its effectiveness in replicating human motion with precision.**

I. Introduction

Nowadays, the development of robots continues to advance rapidly worldwide. Among various types of robots, humanoid robots have drawn significant attention recently due to their ability to interact like humans and perform human tasks. This type of robot requires various training stages of data to ensure it can perform humanoid motions without unexpected behaviours.

To develop a system that is capable of capturing human motion and controlling robots, many researchers came up with different approaches. Filiatrault and Cretu [1] approach this matter by using Kinect for Xbox to capture human arm movements, map them using rotation matrices, and illustrate the system using the Nao robot in real time. Kong et al. [2] developed a real-time IMU-based motion capture system for gait rehabilitation. Zhang et al. [3] use inertial sensors to capture athletes' posture and gait to assess ankle injuries and monitor rehabilitation progress. Meng et al. [4] introduce a motion imitation system that can provide motion to the humanoid robot using human pose estimation in 3D. Zhan and Huang [5] proposed a motion imitation system that extracts 3D coordinates from the human body using a 3D pose estimation model and maps these key points into robots' trajectory files for motion execution. Motamedi et al. [6] investigated the use of visual, vibrotactile, and pressure feedback to control a robotic arm for tasks like grasping and aiming to see if touch could reduce reliance on sight, especially for amputees. Prayudi and Kim [7] developed a cost-effective, high-speed motion capture system for human limbs using a custom IMU module. They use a serial network and addressing alignment for better accuracy through a simple calibration method.

Training humanoid robots to move just like humans is a complicated task when relying only on coding environments and simulation. Limited human motion data on how the robot should behave might lead to some unexpected behaviour in the robot, causing damage to the robot, injury to humans, and damage to the environment. To solve this problem, the data from human motion is introduced to be integrated with the data on the machine, which effectively trains the robot to behave more like a human. By using a wearable controller equipped with multiple Inertial Measurement Units (IMUs), the machine can capture data from the motion of the human character in real life. However, the IMU-based imitation system faces some issues due to gyroscope bias. It can lead to drift, especially on the z-axis, causing interference with motion calculation. To address this issue, the calibration of sensors can optimize the measurement and reduce drift. Besides that, the implementation of a digital low-pass filter (DLPF) and a complementary filter in the data processing also contributes to the drift reduction. Initially, the servo motor-based robots often experience abrupt motion due to the instant movement of the servo from one position to a new one. To solve this issue, we use Spherical Linear Interpolation (SLERP) in motion processing, allowing the servo motor to follow a smooth path and improving the fluidity of the robot. The integration of pulse width modulation (PWM) with SLERP can effectively control the speed of the servo to move smoothly. Additionally, there is another issue with the interactions between users and the robots. Users can't feel the interaction between the robots and the objects during task performance, causing damage to the objects and a lack of tactile feedback. As a solution, the system includes a closed-loop mechanism that provides users with physical feedback from the robots to the controller.

In this paper, we propose a closed-loop system that uses three IMUs to capture human motion, then transmits data via Bluetooth communication and receives haptic feedback from the robot arm simultaneously, ensuring real-time operation. A motion planning mechanism is also introduced in this paper to map the received human motion data and use it to control the robot arm.

II. METHOD

A. System Architecture

The closed-loop system consists of two different embedded systems: one for the controller and another for the 7-DOF robot arm shown in Fig. 1. The controller system, equipped on the subject's arm, transmits movement and pressure data to the robot arm system. The data is used to control actuators while haptic feedback from the gripper is simultaneously sent back to the controller. Both systems utilize an RTOS (real-time operating system) due to its multitasking with priority levels, task scheduling, and resource management. Powered by ESP32, both systems communicate with each other using Bluetooth.

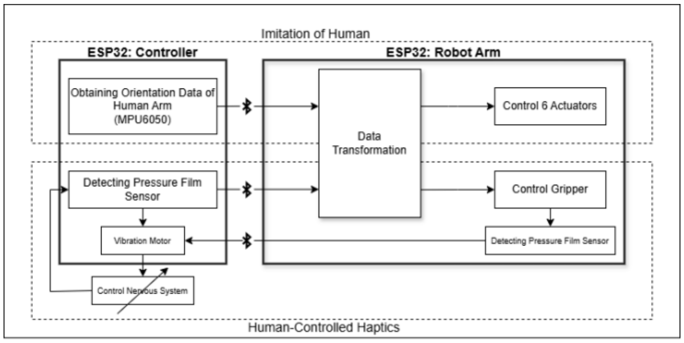


Fig. 1. Diagram of data flow

B. System Design for Controller

The controller utilizes a DLPF (digital low-pass filter) to filter out high-frequency noise from sensor data for smoother and more stable reading. Additionally, the system uses quaternion, a four-dimensional representation of orientation, to avoid issues like gimbal lock. From (1), the quaternion *q* consists of a scalar part *w*, representing the real component of the quaternion, while the vector components *x*, *y*, and *z* define the imaginary part. Additionally, *i*, *j*, and *k* are the unit vectors that define the orientation of the quaternion in space.

$$q = w + xi + yj + zk \quad (1)$$

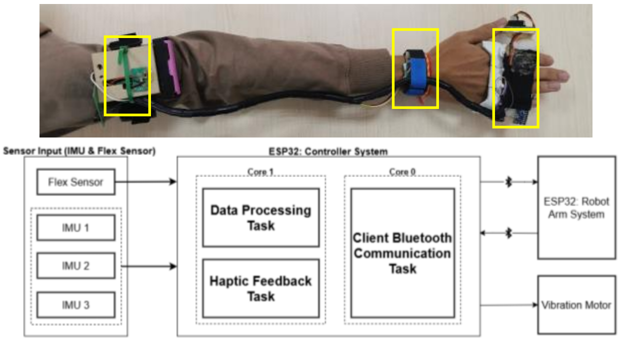


Fig. 2. Controller system diagram with controller representation

In the controller system, RTOS allows the system to operate three important tasks: Client Bluetooth Communication Task (CBCT), Data Processing Task (DPT), and Haptic Feedback Task (HFT) simultaneously. The system operates using ESP32, a microcontroller equipped with dual cores (Core 0 and Core 1), which allows the tasks to operate separately. The DPT and HFT are assigned to Core 1, while CBCT operates on Core 0 with the same level of priority. Each operates with a different stack size to effectively maintain resource management while optimizing its performance.

Before utilizing these tasks, the system needs to go through some essential initialization for the IMU sensor and Bluetooth communication, especially for IMU sensors that require correct bit configuration to read data regarding the accelerometer, gyroscope, and DLPF. Moreover, pre-calibrating each IMU manually allows it to operate with minimal measurement errors.

<u>*Client Bluetooth Communication Task (CBCT):*</u> This task operates on a separate core to ensure no interruption to the communication while performing the other tasks. It handles data transmission and reception between the controller system and the robot arm system. This task helps in data arrangement while maintaining the order of the data in each package according to the predefined head and tail markers.

<u>*Data Processing Task (DPT):*</u> Operating on Core 1 of the ESP32, this task processes extracted data from multiple IMUs and converts it into motion data. In this task, the data were extracted from three different IMUs, as shown in Fig. 2, to capture the movement of the arm by generating different orientations that are crucial for controlling the robot. This project utilizes the MPU6050 as the IMU sensor module. To effectively mimic human movement, those sensors are specifically put in the planned position on the wearable controller to generate essential orientation data for robot arm control, minimizing drift. The processing in this task is as follows:

1. The generated raw data from IMUs (*ax, ay, az, tmp, gx, gy, gz*) are read based on the address and buses to ensure separate operation without any interference.
2. To convert these data to quaternions, it required some processing consisting of a complementary filter (feedback-based sensor fusion), error-state feedback correction, and quaternion integration.
3. The quaternion data is enqueued to the FIFO queue before being transmitted from the controller to the robot via Bluetooth communication.
4. After each package containing movement data is sent out, the controller also receives a haptic feedback package from the server of the robot arm.

<u>*Haptic Feedback Task (FBT):*</u> In this task, the data from the pressure sensor on the gripper of the robot arm was sent back to the controller, resulting in a closed-loop system. It allows users to feel the interaction between the gripper on the controller and the gripper of the robot arm using different levels of vibration or no vibration at all based on the mapped pressure sensor of the gripper.

C. System Design for the Robot Arm

SLERP is used to ensure the smooth orientation transition and consistent angular velocity of the movement of the robot. From (2), $slerp_{(q_1,q_2;u)}$ is the interpolated quaternion, q_1 is the starting quaternion, q_2 is the ending quaternion, *u* is the interpolation factor, and θ is the angle between starting and ending quaternions in quaternion space [8].

$$slerp_{(q_1,q_2;u)} = \frac{\sin(1-u)\theta}{\sin\theta} q_1 + \frac{\sin u\theta}{\sin\theta} q_2 \tag{2}$$

By using RTOS on the ESP32, the Robot Arm system operation is divided into the Server Bluetooth Communication Task (SBCT) and the Motion Data Processing & Control Task (MDPCT), as shown in Fig. 3. The BCT receives motion data from the controller system and transmits feedback from the pressure sensor to the controller system. Additionally, MDPCT processes the received quaternion data and transforms it to new quaternion data using SLERP. To generate the robot's movements, new quaternions from the IMU sensors are processed using quaternion operations such as multiplication, conjugation, normalization, and others. These operations enable the accurate combination of quaternions from multiple IMUs mounted on the subject's arm. This method ensures precise joint rotation by maintaining a parent-child relationship between each segment. Finally, the resulting quaternion values are converted into rotation angles, which are then mapped and constrained to the use of PWM to control the arm's speed and position.

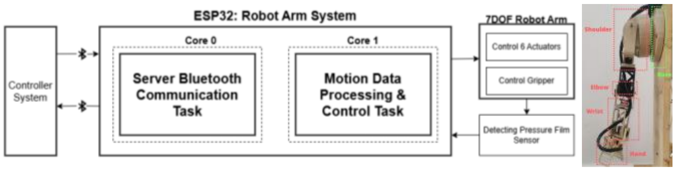


Fig. 3 Robot arm system diagram with robot arm representation

III. Experiment

To evaluate the performance of the robot arm in replicating human arm movements, various experiments were conducted to interpret the movement and control of the robot in real-time control. The experiments were conducted to explore the limitations of the robot arm movement and the reliability of the system in real-time motion imitation. Various movements of the robot arm were tested, including basic movements such as side and forward hand-lifting and raising a hand, and complex movements. The reliability of the communication system is also observed in this section to study the operating range between the operator and the robot arm.

A. Experiment set-up

In the experiments, there are two important hardware parts, which are the controller and the robot arm. The structure of the controller and robot arm was designed using computer-aided design (CAD). We use PLA+ filament to construct the structure of the robot arm and controller due to its strength, printability, and economical price. The controller consists of three IMUs that are capable of capturing movement and orientation. The controller system uses an ESP32 to control those IMUs, a vibration motor, and a flex sensor. In this experiment, the user wears this controller on the right arm with IMUs mounted at a specific position to capture the correct data, as shown in Fig. 4. Initially, the robot arm is composed of seven DOFs, controlled by different models of servo motors, such as RC servo, MG995R, and SG90s, based on different loads and abilities. The robot arm system uses an ESP32 and a PCA9685 to control the PWM of the servo motors.

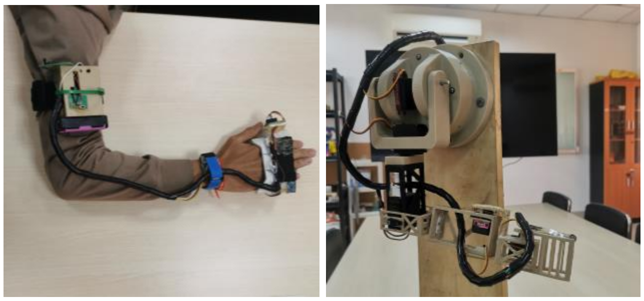


Fig. 4 Controller worn by the subject and robot arm for real-time motion replication

B. Experiment Procedure

In this experiment, we positioned the controller and robot arm below a range of 2 meters to optimize the communication with real-time operation. While the robot arm system functions as a Bluetooth server, the controller system functions as a client that connects to the server using the server address. The design of the experiment is as follows:

1. Firstly, position the IMU sensors on a flat surface for proper calibration, and its value is later used for the process.
2. Secondly, booting up the ESP32 on the robot arm initializes the Bluetooth server by turning on the switch that connects the ESP32 and the battery.
3. Thirdly, power on the controller's ESP32 using the switch, just as was done on the server side, to initialize the Bluetooth client.
4. After that, the ESP32 client will take around 3 to 5 seconds in the process of scanning, discovery, and connecting.
5. Later, connect the power supply to the servo driver in the robot arm system.

C. Experiment Result

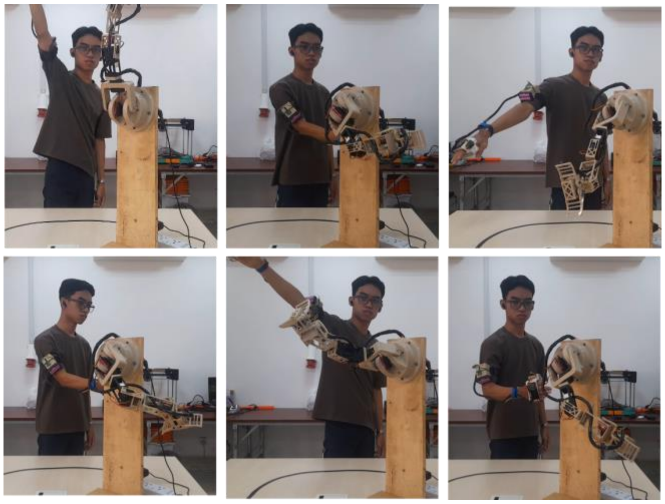


Fig. 5 Sample images showing basic robot movements controlled by the subject

Different experiments in Figs. 5 and 6 showed subjects who controlled the robot using a controller equipped with IMU sensors and moved accordingly to capture both basic and complicated movements. The experiments showed that the robot arm can mimic human motion effectively. It can be observed that the skin of the operator affects the rotation and movement of the robot arm. This occurs because the controller, when mounted on the arm, may not rotate in perfect alignment with the operator's actual arm movement. It is because human bone and skin do not rotate identically, leading to slight deviations in certain robot arm motions. Another impact is the limitation of the actuator, which results in the inability to replicate certain human motions. The actuators used in this method are only capable of supporting rotation within a 180-degree range, while the human arm can rotate beyond this limit. Moreover, a noticeable latency deviation occurs in Bluetooth communication between the robot and the controller once the distance between them increases. Regarding this latency, it is also affected by DLPF configuration in the IMU sensors, which depends on certain bit configurations.

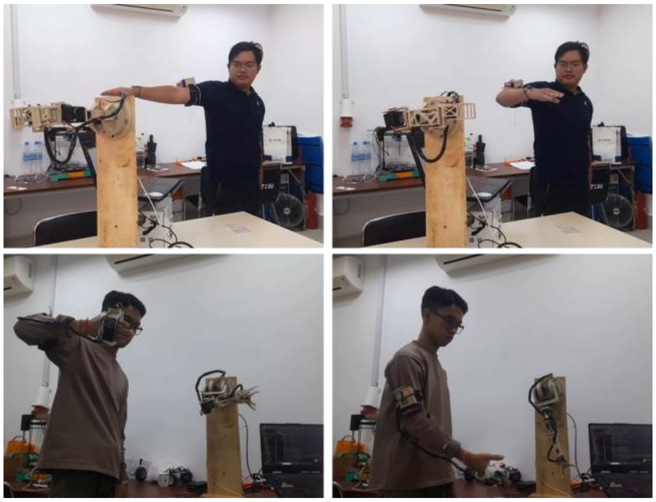


Fig. 6 Sample images illustrating complex robot movements controlled by the subject

IV. Conclusions

In this paper, we propose a closed-loop system that uses three IMUs to capture human motion, then transmits data via Bluetooth communication and receives haptic feedback from the robot arm simultaneously, ensuring real-time operation. Several challenges were encountered, including drifting issues of the IMUs, limitations of data processing and motion mapping, and limited speed of the communication. Despite this, we offer multiple solutions, such as the utilization of DLPF, task handling methods, and quaternions. These solutions allow the robot arm to communicate in real time with the controller on the subject's arm effectively, despite operating on microcontrollers like ESP32. Moreover, data processing techniques allow the robot to operate at an appropriate level with such microcontrollers. In addition, this system can still be improved in the next phase by introducing new components such as a processor, controller, and motion tracking sensors.

References

[1] S. Filiatrault and A. -M. Cretu, "Human arm motion imitation by a humanoid robot," *2014 IEEE International Symposium on Robotic and Sensors Environments (ROSE) Proceedings*, Timisoara, Romania, 2014, pp. 31-36, doi: 10.1109/ROSE.2014.6952979.

[2] W. Kong et al., "Development of a real-time IMU-based motion capture system for gait rehabilitation," *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Shenzhen, China, 2013, pp. 2100-2105, doi: 10.1109/ROBIO.2013.6739779.

[3] S. Zhang, F. Naghdy, D. Stirling, M. Ros, and A. Gray, "Ankle injury assessment using inertial 3D data," *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, Wollongong, NSW, Australia, 2013, pp. 810-815, doi: 10.1109/AIM.2013.6584193.

[4] S. Meng, S. Qiu, T. Liang, and Q. Ren, "Motion Imitation of a Humanoid Robot via Pose Estimation," *2023 35th Chinese Control and Decision Conference (CCDC)*, Yichang, China, 2023, pp. 1526-1532, doi: 10.1109/CCDC58219.2023.10327198.

[5] Z.-F. Zhan and H.-P. Huang, "Imitation System of Humanoid Robots and Its Applications," in *IEEE Open Journal of Circuits and Systems*, vol. 4, pp. 15-24, 2023, doi: 10.1109/OJCAS.2022.3231097.

[6] M. R. Motamedi, J. -B. Chossat, J. -P. Roberge, and V. Duchaine, "Haptic feedback for improved robotic arm control during simple grasp, slippage, and contact detection tasks," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, 2016, pp. 4894-4900, doi: 10.1109/ICRA.2016.7487694.

[7] I. Prayudi and D. Kim, "Design and implementation of IMU-based human arm motion capture system," *2012 IEEE International Conference on Mechatronics and Automation*, Chengdu, China, 2012, pp. 670-675, doi: 10.1109/ICMA.2012.6283221.

[8] Ken Shoemake, "Animating rotation with quaternion curves," *SIGGRAPH Computer Graphics*, vol. 19, no. 3, pp. 245–254, 1985. doi:10.1145/325165.325242.

[9] S.-O. Shin, D. Kim, and Y. -H. Seo, "Controlling Mobile Robot Using IMU and EMG Sensor-Based Gesture Recognition," *2014 Ninth International Conference on Broadband and Wireless Computing, Communication and Applications*, Guangdong, China, 2014, pp. 554-557, doi: 10.1109/BWCCA.2014.145.

[10] C. -J. Lin and H. -Y. Peng, "A study of the human-robot synchronous control based on IMU and EMG sensing of an upper limb," *2022 13th Asian Control Conference (ASCC)*, Jeju, Korea, Republic of, 2022, pp. 1474-1479, doi: 10.23919/ASCC56756.2022.9828042.

[11] R. Fu, Q. Li, S. Wang, and G. Sun, "Teleoperation Method For Controlling Robotic Arm Based On Multi-channel EMG Signals,*" 2023 38th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Hefei, China, 2023, pp. 1264-1268, doi: 10.1109/YAC59482.2023.10401439.

[12] M. Syakir, E. S. Ningrum, and I. Adji Sulistijono, "Teleoperation Robot Arm using Depth Sensor," *2019 International Electronics Symposium (IES)*, Surabaya, Indonesia, 2019, pp. 394-399, doi: 10.1109/ELECSYM.2019.8901679.

[13] U. Muhammad, K. A. Sipra, M. Waqas, and S. Tu, "Applications of Myo Armband Using EMG and IMU Signals," *2020 3rd International Conference on Mechatronics, Robotics and Automation (ICMRA)*, Shanghai, China, 2020, pp. 6-11, doi: 10.1109/ICMRA51221.2020.9398375.

[14] S. Jitpakdeebodin, P. Nararat, P. Duangtoi, K. Eiamsaard, and P. Bamrungthai, "IMU-Based Motion Capture Using Madgwick Filter with 3D Visualization for Robot Teleoperations,*" 2024 9th International Conference on Control and Robotics Engineering (ICCRE)*, Osaka, Japan, 2024, pp. 121-126, doi: 10.1109/ICCRE61448.2024.10589761.

[15] L. Saing, N. Khiev, and K. H. Kim, “Teleoperated haptic robot arm using ESP32 - [Video Demo],” YouTube, Apr. 2025. [Online]. Available: https://youtu.be/SYQrTRivSDM

AI-Powered Child Health Classification: Analyzing VR-Based Eye and Facial Tracking Data Using Deep Learning

*Awais Khan[1], Yongwon Cho[2], Yunyoung Nam[3, *]*

[1]Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Republic of Korea

[2] Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea

[3] Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea

**Contact: ynam@sch.ac.kr*

Abstract— **Understanding child behavior is essential for the early detection of developmental disorders and cognitive impairments. Traditional methods depend on manual observation, which can be inconsistent, time-consuming, and subjective. With enhancement in Artificial Intelligence (AI), Virtual Reality (VR) and automated behavior analysis has appeared as a promising method to improving scalability and accuracy in health monitoring system. Existing behavior recognition methods, lack of real-time monitoring capabilities and fail to provide objective, data-driven visions into a child's cognitive and motor functions. An interactive, immersive and automated system is necessary to accurately distinguish between unhealthy and normal conditions in children using advanced tracking technologies. Objective of this study is to develop a VR-based AI framework using Meta Quest Pro to classify unhealthy and normal child conditions by analyzing eye, and face tracking data. The goal is to provide an automated and accurate solution for child condition assessment. A dataset of 377 features associated to facial expressions, eye gaze, head motion and pupil movement was collected while children performed structured VR-based tasks. After pre-processing, 26 key features were selected using statistical analysis. A hybrid deep learning model was implemented, where Long Short-Term Memory (LSTM) networks processed sequential data, and machine learning classifiers performed classification. Notably, our dataset, obtained from Soonchunhyang University Asan, achieved an average accuracy of 90.9%. This research has significant implications in child healthcare, special education, and assistive technology. Future work includes real-time implementation and large-scale clinical validation.**

I. Introduction

The advancement of virtual reality (VR) and machine learning technologies has opened new avenues for healthcare applications, including the classification of children's health status based on their physical and behavioral responses [1, 2]. Traditional methods of diagnosing developmental and behavioral disorders rely heavily on qualitative assessments, which are often subjective and inconsistent. However, VR-based systems provide an immersive environment where children's movements, eye-tracking data, and facial expressions can be quantitatively analyzed, leading to more reliable health assessments [3].

Repetitive behaviors (RBs) in children with ASD are defined as observable motor actions that occur in a stereotyped or repetitive sequence [4, 5]. These behaviors are often rigid, invariant, and seemingly purposeless, though they may serve as a mechanism for self-regulation in response to changes in routine or unfamiliar stimuli. RBs can be broadly categorized into common and complex behaviors. Common behaviors, such as nail-biting, thumb-sucking, and hair twirling, are frequently observed in both neurotypical and ASD populations, particularly during moments of stress or anxiety [6, 7]. However, complex RBs, including hand-flapping, finger-wiggling, head-spinning, foot-stamping, and body-rocking, are more characteristic of ASD and tend to occur with greater frequency and intensity compared to neurotypical children of the same age [8, 9].

Studies have demonstrated that RBs are highly prevalent among children with ASD, with reported occurrences ranging from 60% to 100% of cases. Given the strong association between repetitive behaviors and ASD, movement-based assessments have emerged as a promising avenue for early diagnosis and classification. Researchers have increasingly explored the potential of using movement features and their frequency as diagnostic biomarkers for ASD. By analyzing motion patterns, it is possible to gain valuable insights into the neurodevelopmental differences between children with and without ASD, enabling more accurate and objective assessments [9].

Virtual reality (VR) technology offers a unique and innovative platform for evaluating movement-based biomarkers in ASD. Unlike traditional assessment methods, VR provides a controlled and immersive environment where children's responses to various stimuli can be systematically monitored and analyzed. With advancements in VR hardware, such as the Meta Quest Pro, it is now possible to capture precise data on head movements, eye-tracking, and facial expressions. These features are critical in distinguishing between healthy and unhealthy children, as they serve as key indicators of neurodevelopmental and behavioural health [10].

A. Problem statement

Accurately assessing a child's health status remains a significant challenge, as traditional methods rely on subjective evaluations and structured observations that often introduce biases and inconsistencies. These conventional approaches may fail to capture subtle movement-based biomarkers that are essential for an objective diagnosis. Additionally, assessments conducted in controlled clinical settings do not always reflect real-world behavioral patterns, limiting their applicability.

Virtual reality (VR) and machine learning offer a promising alternative by enabling real-time, data-driven evaluations in an immersive environment. VR-based assessments provide precise tracking of eye movements, facial expressions, and motor responses, which can be analyzed using machine learning models for accurate classification. However, current studies lack extensive validation due to small sample sizes and limited datasets. Expanding data collection with a larger and more diverse subject pool is necessary to improve the generalizability and effectiveness of automated health assessments.

B. *Major Contributions*

The objective of this study is to address the limitations of current methods by introducing a novel deep learning approach framework for accurate child health status classification. The proposed framework includes the following steps:

- The dataset is first processed to remove unnecessary and irrelevant data, ensuring high-quality input for training. This step includes handling missing values, normalizing data, and filtering out noise to improve model performance.
- After refining the dataset, key features such as movement patterns, facial expressions, and eye-tracking data are extracted. These features are then used to train deep learning models, including Long Short-Term Memory (LSTM) networks.
- The extracted features from trained deep learning models are passed to machine learning classifiers, such as Boosted Trees, Fine KNN, Subspace KNN, and Cubic KNN, to ensure accurate differentiation between healthy and unhealthy subjects.

II. Related work

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by difficulties in social communication, interaction, and the presence of restricted and repetitive patterns of behavior, interests, or activities [11]. It is estimated to affect approximately 1 in 160 children worldwide, with symptoms typically manifesting between the ages of two and four. In some cases, early indicators can be observed as early as six months [12]. While much of the research on ASD has focused on impairments in social interaction, less attention has been given to the repetitive motor behaviors that significantly impact the educational, social, and daily lives of affected individuals [13, 14].

To address these limitations, researchers have turned to technology-driven solutions, such as virtual reality (VR) and machine learning, to improve the accuracy and objectivity of ASD diagnosis. VR-based assessments allow for a controlled yet immersive environment where children's interactions, movements, and behavioral responses can be quantitatively analyzed. Advanced tracking technologies integrated into VR systems can capture head movements, gaze patterns, and facial expressions, providing critical insights into neurodevelopmental differences. Unlike traditional assessments, VR enables real-time data collection and eliminates the need for subjective scoring [15, 16].

Recent advances in cognitive neuroscience have also contributed to the development of implicit assessment techniques for ASD. Implicit measures capture automatic biological responses that occur outside conscious awareness, offering a more objective means of understanding social interactions and behavioral patterns. Biomarkers such as electrodermal activity (EDA), functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), electroencephalography (EEG), eye tracking, and heart rate variability (HRT) have been explored in ASD research. For example, fMRI studies have shown that ASD is associated with hyperactivity in neural circuits, particularly in the cingulate posterior cortex and portions of the insula. EEG research has further suggested that individuals with ASD exhibit altered neural activity, particularly in social contexts [17, 18].

The growing field of VR-based ASD assessment, combined with machine learning and advanced tracking technologies, holds significant potential in addressing the limitations of traditional diagnostic approaches. By leveraging these technologies, researchers can develop more objective and reliable methods for classifying ASD and distinguishing between healthy and unhealthy children based on movement-related biomarkers. Future studies should focus on refining these methodologies, expanding their applicability across diverse populations, and ensuring ethical considerations in pediatric assessments. The integration of AI-driven diagnostics and immersive VR environments represents a promising step forward in improving ASD diagnosis and early intervention strategies [19, 20].

III. Proposed Method

In this section, we propose a new deep learning method for child health classification as illustrated in Figure 1. The methodology includes several distinct stages: initial pre-processing of the data, extraction of features utilizing LSTM and RNN models, and at last classification. This approach employs advanced techniques in deep learning to enhance the performance of LSTM deep learning model. Subsequent to the extraction of features from these models, using the VR based eye and facial tracking data, the feature vectors derived from the eye and facial data via both models are passed to the machine learning classifiers for the final classification. After feature extraction, the feature vectors derived from the eye-tracking and facial tracking data are processed using machine learning classifiers to achieve optimal classification performance. Specifically, we employ Boosted Tree, Fine K-Nearest Neighbors (KNN), Subspace KNN, and Cubic KNN classifiers to analyze the extracted features and classify the children's health status. This integration of deep learning and machine learning techniques provides a more robust and accurate classification system, ensuring reliable differentiation between healthy and unhealthy children based on VR-based eye and facial tracking data.

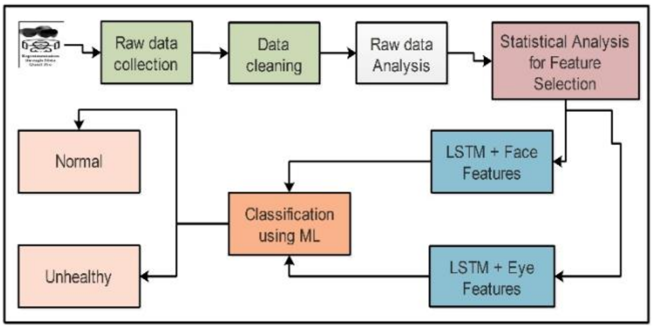


Figure 1. Proposed diagram of child health classification.

A. *Data Collection*

This study included a sample of 112 children between the ages of 4 and 7 years, consisting of 54 children diagnosed as unhealthy and 58 with typical healthy development. The classification of health status was determined by medical professionals. The dataset was collected at the CRC center in Soonchunhyang University, Asan, South Korea. A VR-based learning environment was developed to assess the normal and unhealthy behaviors of children. The children were instructed to wear the Meta Quest Pro headset and perform tasks within the VR environment under the guidance of a teacher. Before engaging in the actual tasks, participants were given adequate training to familiarize themselves with the Meta Quest Pro and the virtual environment.

The children were required to complete three specific tasks designed to assess their cognitive and motor abilities. The first task involved sorting balls of different colors (red, green, yellow, and blue) into their corresponding boxes. The second task required placing objects such as cars, animals, and fish into their designated categories. The third task involved following the instructions given by a VR-based teacher to correctly place an object into a specified box. These tasks were carefully designed to evaluate behavioral responses, motor coordination, and cognitive processing in both healthy and unhealthy children. The use of VR in this study ensures a controlled yet engaging environment where children's actions can be monitored objectively. Extracted eyes and facial features from Meta Quest Pro are listed in table 1. Environment setup for data collection is shown in Figure. 2.

Table 1. Eye and facial features extracted from Meta Quest Pro.

Eye Features	Face Features
eyes_look_down_l	inner_brow_raiser_l
eyes_look_down_r	inner_brow_raiser_r
eyes_look_left_l	cheek_raiser_l
eyes_look_left_r	cheek_raiser_r
L_eye_cls	lid_tightener_l
R_eye_cls	lid_tightener_r
l_eye_pnt_x	Outer_brow_raiser_r
l_eye_pnt_y'	Outer_brow_raiser_l
'r_eye_pnt_x	Upper_lid_raiser_r
'r_eye_pnt_y	Upper_lid_raiser_l
l_eye_pnt_x	Upper_lip_raiser_l
l_eye_pnt_y'	Upper_lip_raiser_r
	Lip_corner_puller_r

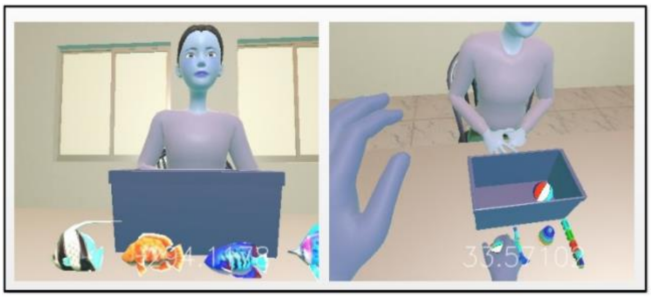


Figure 2. Environment set for dataset collection.

IV. RESULTS AND DISCUSSION

In this section, we have presented the detailed experimental results of the proposed framework. The results are presented using visual graphs and well-defined performance measures to provide a comprehensive and clear valuation of our methodology's. In this study, the dataset was divided into training and testing sets with ratio of 70:30. The training process was configured with specific parameters, including 100 iterations, 100 epochs, a minibatch size of 34, and a learning rate set at 0.0001. Stochastic Gradient Descent (SGD) served as the optimization algorithm. A 5-fold cross-validation was executed, assessing multiple classifiers across a range of performance metrics, including precision, rate, recall, and accuracy. All simulations were conducted using MATLAB 2024a.

A. *Experimental results for face features using LSTM*

In this subsection, the classification results of child health condition detection have been discussed, employing LSTM deep-learning models. The dataset used for experiments was collected using VR-based motion tracking, capturing critical behavioral and movement-related features. Various ML classifiers were applied, including Boosted Tree, Fine K-Nearest Neighbors (KNN), Subspace KNN, and Cubic KNN. The results for the LSTM model with VR-based eye feature data is presented in Table 2. The highest accuracy of 90.2% was achieved by Boosted Tree classifier when using the LSTM model with the computational time, recall rate, precision rate, and AUC values of 2.6 s, 89.6%, 92.3%, and 0.96, respectively. The second-highest accuracy of 83.8% was obtained with the FKNN classifier, resulting in computational time, recall rate, precision rate, and AUC values of 2.1 s, 78.4%, 79.9%, and 0.78. The difference in accuracy between the top two models was only 6.4%. The third-best accuracy, standing at 72.7%, was achieved by the subspace KNN classifier, along with computational time, recall rate, precision rate, and AUC values of 0.46 s, 64.9%, 60.4%, and 0.71, respectively. The confusion matrix eye feature using LSTM and boosted tree is shown in Figure. 3.

Table 2. Experimental results for face features using LSTM.

ML models	Accuracy	Time	Precision rate	Recall rate	AUC
Boosted Tree	90.2%	2.6	92.3	89.6	0.96

Fine KNN	83.8%	2.1	79.9	78.4	0.78
Subspace KNN	72.7%	0.46	60.4	64.9	0.71
Cubic KNN	70.6	2.2	68.8	69.4	0.68

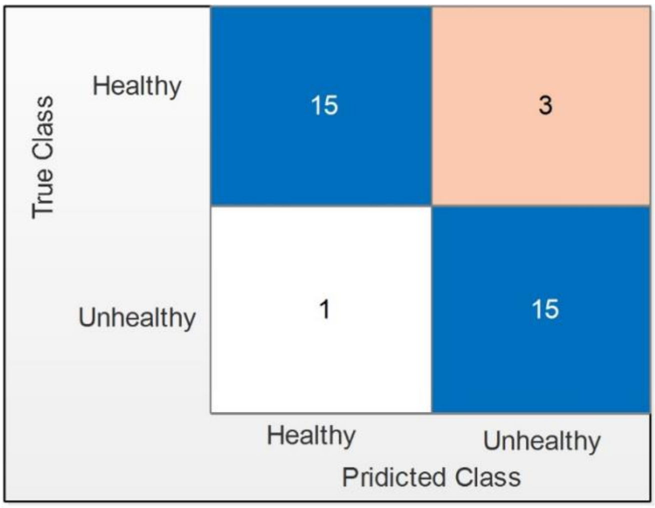


Figure 3. Confusion matrix for eye features using LSTM and Boosted tree.

B. *Experimental results for eye features using LSTM*

In this subsection, the classification results LSTM deep-learning models have been discussed. The dataset used for experiments was collected using VR-based motion tracking, capturing critical behavioral and movement-related features. Various ML classifiers were applied, including Boosted Tree, Fine K-Nearest Neighbors (KNN), Subspace KNN, and Cubic KNN. The results for the LSTM model with VR-based eye feature data is presented in Table 3. The highest accuracy of 87.4% was achieved by Boosted Tree classifier when using the LSTM model. This model yielded computational time, recall rate, precision rate, and AUC values of 2.9 s, 90.4%, 89.1%, and 0.89, respectively. The second-highest accuracy of 83.3% was obtained with the Subspace KNN classifier, resulting in computational time, recall rate, precision rate, and AUC values of 0.4 s, 78.8%, 79.8%, and 0.78. The difference in accuracy between the top two models was only 4.1%. The confusion matrix for LSTM and bossted tree is shown in Figure. 4.

Table 3. Experimental results for eye features using LSTM.

ML models	Accuracy	Time	Precision rate	Recall rate	AUC
Boosted Tree	87.4%	2.9	89.1	90.4	0.89
Fine KNN	70.9%	1.7	69.1	6.4	0.68
Subspace KNN	83.3%	0.4	79.8	78.8	0.78
Cubic KNN	71.4	3.7	70.9	69.5	0.72

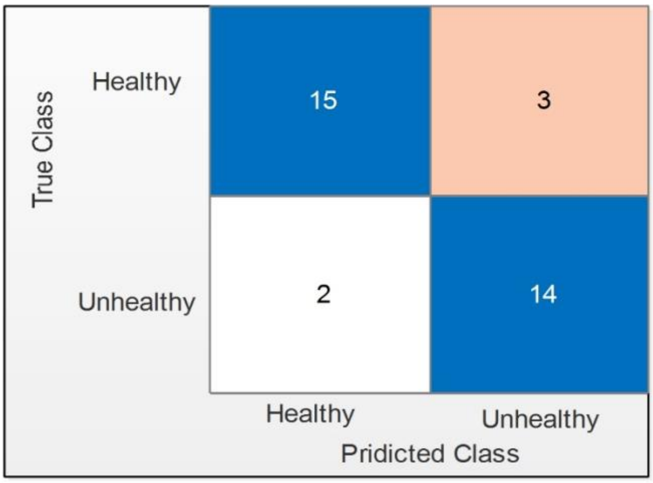


Figure 4. Confusion matrix for eye features using LSTM and Boosted tree.

V. Conclusion

This study highlights the effectiveness of integrating VR-based assessments and machine learning models for child health condition classification, overcoming limitations of traditional diagnostic methods. By analyzing movement-related biomarkers, our approach enhances objectivity and reliability. The results demonstrate high classification accuracy, reinforcing the potential of advanced tracking technologies. Comparison with existing work is presented in Table 4. However, to improve generalizability, future work should include a larger dataset with more diverse subjects. Expanding the feature set and exploring real-time clinical applications will further enhance the effectiveness of this approach in ASD diagnosis.

Table 4. Comparison with existing methods.

Reference	Method	Accuracy	Year
[21]	ML on VR-captured body movement data	85%	2020
[22]	ML on behavioural data	80%	2018
Proposed	LSTM, ML and VR data	90.9%	2025

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176).

Reference

[1] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®); American Psychiatric Pub: Washington, DC, USA, 2013.

[2] World Health Organization. Available online: https://www.who.int/news-room/fact

sheets/detail/autismspectrum-disorders (accessed on 20 November 2019).

[3] Anagnostou, E.; Zwaigenbaum, L.; Szatmari, P.; Fombonne, E.; Fernandez, B.A.; Woodbury-Smith, M.; Buchanan, J.A. Autism spectrum disorder: Advances in evidence-based practice. Cmaj 2014, 186, 509–519.

[4] Lord, C.; Risi, S.; DiLavore, P.S.; Shulman, C.; Thurm, A.; Pickles, A. Autism from 2 to 9 years of age. Arch. Gen. Psychiatry 2006, 63, 694–701.

[5] Schmidt, L.; Kirchner, J.; Strunz, S.; Bro ´zus, J.; Ritter, K.; Roepke, S.; Dziobek, I. Psychosocial functioning and life satisfaction in adults with autism spectrum disorder without intellectual impairment. J. Clin. Psychol. 2015, 71, 1259–1268.

[6] Turner, M. Annotation: Repetitive behaviour in autism: A review of psychological research. J. Child Psychol. Psychiatry Allied Discip. 1999, 40, 839–849.

[7] Lewis, M.H.; Bodfish, J.W. Repetitive behavior disorders in autism. Ment. Retard. Dev. Disabil. Res. Rev. 1998, 4, 80–89.

[8] Ghanizadeh, A. Clinical approach to motor stereotypies in autistic children. Iran. J. Pediatr. 2010, 20, 149.

[9] Mahone, E.M.; Bridges, D.; Prahme, C.; Singer, H.S. Repetitive arm and hand movements (complex motor

[10] stereotypies) in children. J. Pediatr. 2004, 145, 391–395.

[11] MacDonald, R.; Green, G.; Mansfield, R.; Geckeler, A.; Gardenier, N.; Anderson, J.; Sanchez, J. Stereotypy in young children with autism and typically developing children. Res. Dev. Disabil. 2007, 28, 266–277.

[12] Singer, H.S. Motor stereotypies. Semin. Pediatr. Neurol. 2009, 16, 77–81.

[13] Lidstone, J.; Uljarevi´c, M.; Sullivan, J.; Rodgers, J.; McConachie, H.; Freeston, M.; Leekam, S. Relations among restricted and repetitive behaviors, anxiety and sensory features in children with autism spectrum disorders. Res. Autism Spectr. Disord. 2014, 8, 82–92.

[14] Bodfish, J.W.; Symons, F.J.; Parker, D.E.; Lewis, M.H. Varieties of repetitive behavior in autism: Comparisons to mental retardation. J. Autism Dev. Disord. 2000, 30, 237–243. [CrossRef] [PubMed]

[15] Campbell, M.; Locascio, J.J.; Choroco, M.C.; Spencer, E.K.; Malone, R.P.; Kafantaris, V.; Overall, J.E. Stereotypies and tardive dyskinesia: Abnormal movements in autistic children. Psychopharmacol. Bull. 1990, 26, 260–266.

[16] Goldman, S.; Wang, C.; Salgado, M.W.; Greene, P.E.; Kim, M.; Rapin, I. Motor stereotypies in children with autism and other developmental disorders. Dev. Med. Child Neurol. 2009, 51, 30–38.

[17] Lord, C.; Rutter, M.; DiLavore, P.C.; Risi, S.A. Diagnostic Observation Schedule-WPS (ADOS-WPS); Western Psychological Services: Los Angeles, CA, USA, 1999.

[18] Lord, C.; Rutter, M.; Le Couteur, A. Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J. Autism Dev. Disord. 1994, 24, 659–685.

[19] Goldstein, S.; Naglieri, J.A.; Ozonoff, S. Assessment of Autism Spectrum Disorder; The Guilford Press: New York, NY, USA, 2009.

[20] Gonçalves, N.; Rodrigues, J.L.; Costa, S.; Soares, F. Preliminary study on determining stereotypical motor movements. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 1598–1601.

[21] Alcaniz Raya, M., Marín-Morales, J., Minissi, M. E., Teruel Garcia, G., Abad, L., & Chicchi Giglioli, I. A. (2020). Machine learning and virtual reality on body movements' behaviors to classify children with autism spectrum disorder. Journal of clinical medicine, 9(5), 1260.

[22] Porayska-Pomsta, K., Alcorn, A. M., Avramides, K., Beale, S., Bernardini, S., Foster, M. E., ... & Smith, T. J. (2018). Blending human and artificial intelligence to support autistic children's social communication skills. ACM Transactions on Computer-Human Interaction (TOCHI), 25(6), 1-35.

A Multimodal Emotion Recognition Model Incorporating Arousal and Valence with Incomplete Data

Y. Shin[1], B. Kim[2], Y. Kang[2], and S. Seo[1,*]
[1]*School of Art and Technology, Chung-Ang University, Anseong, 17546, South Korea*
[2]*Department of Applied Art and Technology, Chung-Ang University, Anseong, 17546, South Korea*
*Contact: sanghyun@cau.ac.kr, phone +82-10 7273 0318

Abstract— **The advent of artificial intelligence has turned emotion recognition into a pivotal research topic. However, conventional categorical emotion classification models often struggle to capture the complexity of human emotions. While valence-arousal emotion models overcome this limitation by representing emotions on continuous axes, they have been predominantly applied to visual data, limiting their effectiveness across other modalities. Moreover, in real-world scenarios with incomplete data, the performance of conventional multimodal models can degrade significantly. In this paper, we propose a robust multimodal emotion recognition model that predicts arousal and valence while maintaining high performance even with incomplete data. Our experiments demonstrate that combining an audio-based arousal predictor with a text-based valence predictor yields the best performance in both regression metrics and categorical emotion classification accuracy. These results indicate that incorporating arousal and valence predictions is effective for enhancing emotion recognition performance and our approach remains robust even when parts of the multimodal data are missing.**

I. Introduction

In recent years, emotion recognition has emerged as a prominent research challenge with the rapid advancement of artificial intelligence. In particular, multimodal emotion recognition — which integrates audio, text, facial expressions, and biometric signals — has received growing attention as an effective way to analyze the complexity of human emotion. However, conventional categorical emotion models have limitations when describing complex and nuanced human emotions [1]. This limitation has led to increased focus on continuous two-dimensional methods such as Russell's emotion model [2]. Russell’s emotion model uses arousal and valence to represent complex emotions, but it has been applied mainly to specific data modalities. Additionally, traditional multimodal models are generally developed under the assumption that all modalities are always available, which can result in significant performance degradation under incomplete data conditions [3]. Moreover, the nuances of the Korean language allow the same expression to convey different meanings depending on the context or vocal intensity.

To solve these issues, we propose a multimodal emotion recognition model that incorporates arousal and valence. The model is designed to maintain stable performance even in incomplete data situations as demonstrated using the KEMDy20 dataset [4]. This approach improves the accuracy and robustness of multimodal emotion recognition and demonstrates the feasibility of effectively adapting arousal and valence to a wider range of data formats.

II. Related work

A. Arousal-valence model

In contrast to simple categorical classifications, the Arousal-Valence model represents human emotions as continuous dimensions. Russell’s model places arousal on the vertical axis and valence on the horizontal axis, visualizing various emotional states in a coordinate form [5]. Arousal indicates the level of activation, with lower values suggesting a calm and stable state, whereas higher values denote an excited and energized state. Valence reflects the degree of positivity or negativity of an emotion. By displaying the relative distances among different emotional states, Russell’s model overcomes the limitations of category-based approaches and flexibly captures blended or nuanced emotions.

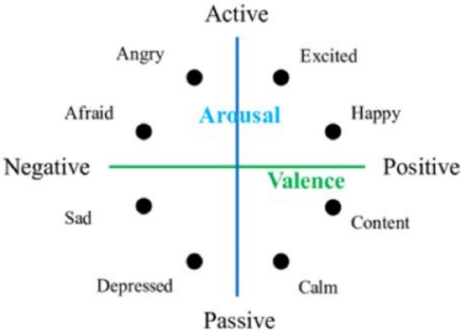
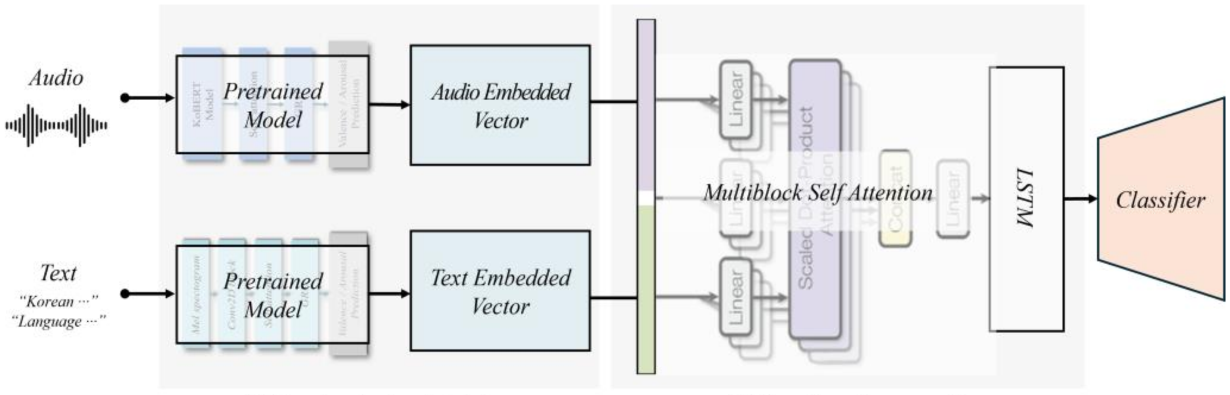


Fig. 1 Emotion examples on an Arousal-valence model

B. Incomplete Data Situation

Multimodal emotion recognition has been shown to outperform single-modality approaches by utilizing information from diverse sources. However, in real-world environments, some modalities may be absent due to recording errors or communication delays. Most models assume the presence of all modalities, making them less robust to missing data. As a result, these models may exhibit reduced performance when data is incomplete. To address this challenge, learning techniques for multimodal missing data restoration have been proposed to compensate for partially corrupted data.



(a) Step of extracting data feature

(b) Step of emotion recognition

Fig. 2 Pipeline of the multimodal emotion recognition model

III. Methodology

This section provides a comprehensive overview of the proposed model for multimodal emotion recognition, including its structural framework and the training process. Fig. 2 presents the complete framework, which jointly models arousal and valence. Fig.2 (a) depicts the modality-specific regression stage: raw speech and its transcript are processed by an audio encoder and a text encoder, respectively, each optimized to predict continuous arousal and valence scores. The resulting time-series embeddings are then concatenated, as illustrated in Fig. 2 (b), and passed through a fusion network that classifies the input into five discrete emotion categories. The model is trained end-to-end, with the regression objectives acting as auxiliary tasks that regularize the final classifier.

A. Audio-based prediction model and feature extraction

The audio data used as input to the audio encoder was resampled at a rate of 16,000 Hz and transformed into a Mel spectrogram to extract features in the time and frequency domains. After obtaining the spectrum using a Short-Time Fourier Transform (STFT), a two-dimensional convolutional neural network (Conv2D) block and a transformer-based self-attention block are iteratively applied to the Mel spectrogram, which is generated using a Mel-filter bank [6,7]. The process extracts neighboring frequency band characteristics and temporal context concurrently.

The audio encoder incorporates temporal sequence information through a gated recurrent unit (GRU) layer, which is a type of recurrent neural network (RNN). The time-series embedding obtained from the Conv2D and self-attention blocks is passed to the GRU to produce a hidden state. The hidden state at the final step is fed into a fully connected layer (FCL) to predict arousal and valence.

B. Text-based prediction model and feature extraction

The text encoder employed the KoBERT model, a pre-trained language model trained on a substantial Korean dataset [8]. Most of the parameters of the KoBERT model are fixed to preserve the pre-trained language understanding capability, while only the parameters of the upper layers are updated to focus on learning arousal and valence. A self-attention block is applied to the output sequence of KoBERT, assigning greater weights to contextually important words in the embedding space and extracting text features directly associated with emotions. The text encoder also integrates information from the output sequence through a GRU layer, predicting arousal and valence using the final hidden state. Unlike the audio encoder, the text encoder reflects sequential information through KoBERT encoding and enhances contextual understanding through a self-attention block. Consequently, the GRU layer is relatively straightforward and primarily functions to summarize the overall sentence embedding.

C. Emotion Recognition

The emotion embeddings produced by each modality encoder are integrated for the final emotion classification during the multimodal combination process. The time series embedding sequences generated by the audio encoder and text encoder are concatenated along the sequence dimension to form a single multimodal sequence. This fused sequence is then passed through a multi-head attention block, which learns cross-modal correlations between the audio and text embeddings. The resulting integrated sequence, combining both modalities, is fed into a Long Short-Term Memory (LSTM) network as input to the final emotion classifier. The LSTM models the sequential patterns of the integrated sequence to find the last hidden state which serves as a comprehensive representation of the utterance's emotional content. This vector is then passed through a fully connected layer to map it to an emotion class. The final classification output is transformed into a probability through a softmax function, and the entire model is trained to minimize cross-entropy loss.

D. Handling Incomplete Data Scenarios

In order to address scenarios involving incomplete data, independent encoder pathways for the audio and text modalities are preserved. This enables the model to operate when either modality is absent. The model is trained to simulate missing-modality conditions through a process known as modality-drop data augmentation, which serves to mitigate performance degradation during inference. Robustness is further enhanced by alternately freezing and unfreezing the parameters of the pre-trained encoders. For each training instance, either the audio or text stream is randomly omitted with a predefined probability, compelling the network to predict arousal and valence solely from the remaining modality.

IV. Experiments

A. Data preprocessing

In this paper, we employed KEMDy20, a Korean multimodal emotion dataset released by the Electronics and Telecommunications Research Institute (ETRI) in 2020. The original dataset exhibited a pronounced bias toward neutral emotions and contained multiple labels. To address this, we implemented a preprocessing step to adjust the class distribution. Furthermore, emotion classes with exceedingly low sample counts were removed, as certain emotions possessed data distributions of such sparsity that model training became arduous. Specifically, utterances were excluded from the dataset if they had more than one emotion label or if the number of samples for a particular emotion was less than 100. This resulted in a final set of five emotions: neutral, happy, sad, angry, and surprise. To mitigate the effects of over-sampling, we down-sampled the data in each emotion class to a maximum of 600. For labels with an insufficient number of samples, data augmentation was applied to balance the data distribution. For audio data, a technique that masks some frequency-time bins was employed, and for text data, the data was augmented by masking random words. The change in label distribution after each preprocessing step is illustrated in Fig. 3. It is evident from this graph that the class imbalance problem is mitigated.

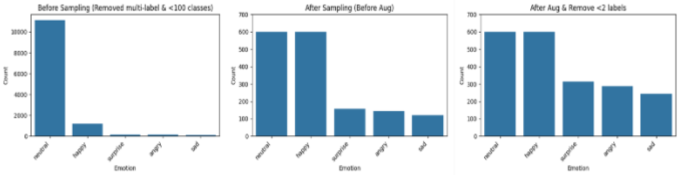


Fig. 3 Visualization of label distribution

B. Experimental setup

We constructed four models with different characteristics to perform comparative experiments. All models use the same preprocessed data and the same initial parameters, with structural differences based on which modalities of arousal and valence are utilized for prediction.

Four models were developed for our experiments, and Table 1 presents an overview of these configurations. All models use the same dataset and initial parameters. However, their architecture differs based on which modalities are used to predict arousal and valence.

TABLE I
Structure of the Models

Model number	Arousal prediction model	Valence prediction model
Model 1	audio-based	text-based
Model 2	text-based	audio-based
Model 3	audio-based	audio-based
Model 4	text-based	text-based

C. Experimental results

Table 2 presents three performance metrics for arousal and valence prediction across the four models. The audio-based arousal predictor records the lowest mean absolute error (MAE) and the highest concordance correlation coefficient (CCC), indicating that prosodic features such as pitch and intensity are highly informative for arousal estimation. In contrast, the text-based valence predictor effectively captures contextual polarity, yet its greater prediction dispersion yields a moderately higher MAE. Although the audio-based valence predictor and the text-based arousal predictor also maintain low MAE values, their reduced CCCs imply that a low error alone is insufficient to guarantee overall agreement.

Table 3 reports the accuracy, precision, recall, and F1-score obtained by the four candidate models. Model 1 delivers the best overall performance, achieving an accuracy of 0.7108 and an F1-score of 0.7637. Both multimodal systems (Models 1 and 2) outperform the unimodal (Model 3: text-only, Model 4: audio-only). Model 2 attains an F1-score of 0.758—only slightly below that of Model 1—but its higher precision and lower recall suggest a bias toward a subset of emotion classes. In contrast, Model 3 records an accuracy of 0.5289 and an F1-score of 0.6314, underscoring the difficulty of classifying emotions from audio alone when both arousal and valence must be inferred without lexical context. Model 4 shows comparable performance (accuracy = 0.5344, F1-score = 0.6160); most of its errors arise from failing to capture prosodic differences between neutral and non-neutral states, leading to confusions between high-energy and low-energy emotions.

TABLE 2
Arousal and Valence Regression Metrics

Model	Arousal			Valence		
	MAE	RMSE	CCC	MAE	RMSE	CCC
Model 1	0.24	0.30	0.48	0.47	0.60	0.34
Model 2	0.27	0.35	0.29	0.36	0.45	0.28
Model 3	0.24	0.30	0.48	0.36	0.45	0.28
Model 4	0.27	0.35	0.29	0.47	0.60	0.34

TABLE 3
Emotion Classification Metrics

Model	Accuracy	Precision	Recall	F1-score
Model 1	0.71	0.87	0.71	0.76
Model 2	0.68	0.89	0.68	0.76
Model 3	0.53	0.85	0.53	0.63
Model 4	0.53	0.87	0.53	0.62

Fig. 4 shows the confusion matrices for each model, illustrating cases of misclassification during emotion recognition. All four models exhibit some confusion between happiness and neutral, likely because the content and intonation of "bright neutral" speech may overlap with mildly happy utterances. Most models also tend to misclassify various emotions as neutral, possibly influenced by the larger proportion of neutral data. Additionally, this overclassification suggests a prediction bias toward neutral when the model encounters ambiguous or complex emotional states.

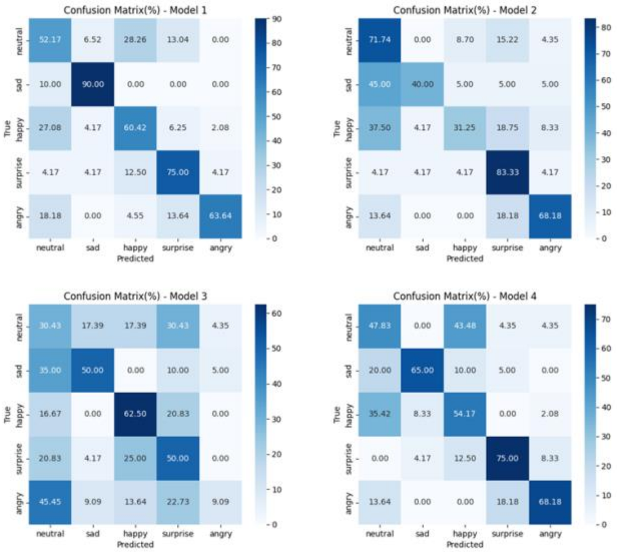


Fig. 4 Confusion Matrix for Each Model

V. Conclusion

In this paper, we present a multimodal emotion recognition model based on arousal and valence and a categorical classification of five emotions (happy, sad, angry, surprise, and neutral). Independent encoders are built for audio and text, respectively, and trained to predict arousal and valence according to each modality. These embeddings and regression outputs are then fused in a multimodal integration module for final emotion classification. To ensure robustness even under incomplete data conditions that frequently occur in real-world settings, we employ a partial parameter-freezing strategy and introduce data augmentation techniques.

Experiments on the Korean dialog dataset KEMDy20 using various model configurations indicate that Model 1 achieves the most balanced performance across arousal and valence regression metrics and emotion classification accuracy. This finding suggests that combining audio-based arousal prediction with text-based valence prediction can effectively enhance emotion classification.

In future work, we plan to expand the scope of multimodal emotion recognition by incorporating visual data, such as facial expressions, as well as biometric signals, and by designing an architecture that dynamically learns the relative importance of each modality. We anticipate that these extensions will facilitate the development of a robust Korean multimodal emotion recognition model applicable to diverse domains, such as conversational AI, dialog analysis tools, and emotion-driven recommendation systems.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176, Contribution Rate: 50%) and by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program(IITP-2025-RS-2023-00266605) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) (Contribution Rate: 50%)

References

[1] P. Ekman, "Are there basic emotions?" 1992.

[2] S. Anders *et al.*, “Brain activity underlying emotional valence and arousal: a response-related fMRI study,” *Human Brain Mapping*, vol. 23, no. 4, pp. 200–209, 2004. doi:10.1002/hbm.20048

[3] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, “Deep adversarial learning for multi-modality missing data completion,” in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, 2018, pp. 1158–1166.

[4] K. J. Noh and H. Jeong, “KEMDy20” [Online]. Available: https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR

[5] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[6] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” *arXiv preprint* arXiv:1511.08458, 2015.

[7] A. Vaswani, “Attention is all you need,” *arXiv preprint* arXiv:1706.03762, 2017.

[8] SKTBrain, “KoBERT” [Online]. Available: https://github.com/SKT-Brain/Kob

Enhancing Multi-Label Emotion Recognition with a Multimodal Deep Learning Approach

Gati L. Martin[1], Jiyoung Woo[2*], and Yunyoung Nam[3]
[1,2]*Department of Future Convergence Technology, Soonchunhyang University, Asan 31538, Republic of Korea*
[3]*Department of ICT Convergence, Soonchunhyang University, Asan 31538, Republic of Korea*
*Contact: jywoo@sch.ac.kr

Abstract—**Emotion detection is a complex task due to variations in acoustic features influenced by factors such as age, gender, culture, language, and developmental stages. Many studies have mainly focused on a narrow range of basic emotions, whereas real-world emotional expressions are often more complex and nuanced. Natural conversations usually involve additional emotions, which can appear in overlapping or blended forms. Recognizing emotions in speech, especially in children, is crucial for applications like mental health monitoring, online education, and the early detection of emotional or developmental challenges. A multimodal deep learning framework is proposed in this study that integrates acoustic, visual, and textual features to improve child emotion detection. A Deep Q-Network-based reinforcement learning feature selection was applied to the acoustic modality, leading to a 2.02% performance improvement over the baseline model. The proposed multimodal approach outperforms traditional unimodal approaches, achieving a weighted F1 score of 66.15%. Experimental results demonstrate that incorporating multiple modalities significantly enhances performance compared to using a single modality. This highlights the complementary strengths of each modality in understanding and capturing the subtle and complex nature of emotional expressions.**

I. Introduction

Emotions are a vital part of language, and they are well-known for being nuanced and complex. Speech emotion recognition (SER) has been shown to be valuable in various applications that involve human-machine interaction, such as therapist diagnostic tools, intelligent voice assistants, car dashboard systems, and online learning platforms [1]. Emotion detection through vocal expression has been a focus across multiple disciplines, including psychology, linguistics, and engineering. Although many studies have primarily focused on six fundamental emotions [2]: fear, happiness, anger, surprise, and sadness, real-world emotional expressions are often more nuanced [3]. Natural conversation exhibits additional emotions like depression, excitement, curiosity, and frustration, which usually appear in overlapping or blended forms. For instance, in the statement, "*This is fascinating! Could you show more examples to help illustrate the concept?*". The statement reflects both curiosity (this is fascinating!) and attentiveness ("Could you show more examples?), which can be helpful for educators to improve online learning experiences.

Emotion detection is complex due to variations in acoustic features influenced by age, gender, culture, language, and developmental stages. Prosodic features such as pitch, loudness, and rhythm are particularly correlated with these factors, making them key indicators of emotional states. For instance, [4] reports accuracy differences of 93.3% for males, 89.4% for females, and 83.3% for children, demonstrating the impact of these variations. Emotion can be recognized through multiple cues, including speech, facial expressions, and physiological signals. Among these, speech and facial expressions are the most natural and effective for conveying emotions. A study by [5] found that in audiovisual emotion recognition, facial expressions contributed 55% to emotional content, voice 38%, and text only 7%. This suggests that while speech-based recognition is effective, integrating additional modalities such as video, images, or text—common in real-world scenarios—can enhance SER performance.

Several studies have studied multimodal interconnections of acoustic, textual, audio, and visual data [6-8]. Central to this has been the design of effective network architectures and loss functions. Deep learning models often optimize performance by adjusting certain loss functions [9]. This flexibility enables deep learning methods to effectively model interactions between different modalities, extracting emotionally meaningful features for more accurate emotion recognition.

In the study [6], the authors introduced a neural network that enhances the utilization of temporal relationships between audio and video modalities in cross-modal fusion through attention mechanisms. By employing a 3D CNN for video processing and a 1D CNN for audio, they achieved an accuracy of 49.2% on a custom dataset of Russian-speaking children aged 5 to 11. [7] presented the multimodal database with facial expressions, body gestures, voice, and physiological signals. They explored the use of deep learning architectures (DBNs and CDBNs) for multimodal emotion recognition, achieving an accuracy of over 80%. Authors of [8] proposed a transformer-based approach, Self-supervised Multi-Label Peer Collaborative Distillation (SeMuL-PCD), integrating a multimodal distillation loss and a self-supervised contrastive objective to enhance generalization across demographics. Evaluations on three datasets, including one focused on children, showed that leveraging audio-video features achieved 91.24% accuracy, outperforming video-only

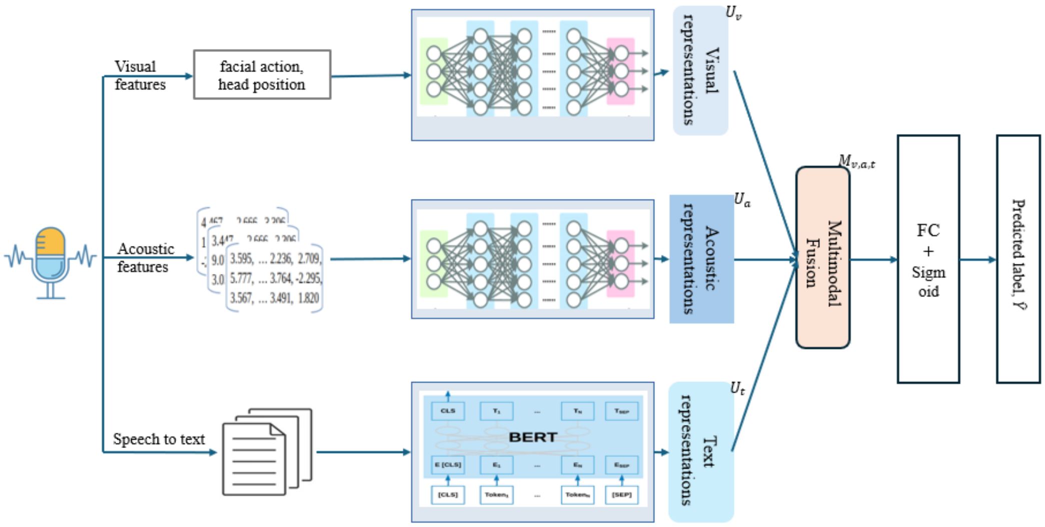


Fig. 1 An overview of proposed model

and audio-only modalities by 4.0% and 5.21%, respectively. The authors of [10] utilize a CNN for facial expressions and a VGG-based feature extractor for audio, where audio is represented as mel-spectrogram images. Their method has shown strong performance in capturing effective representations of facial expressions.

Our study aims to contribute to the growing research in emotion recognition, particularly in children's speech. Children may struggle with pronunciation, vocabulary, or maintaining contextual coherence in textual data, making it difficult to extract precise emotional cues from speech transcripts alone [11]. Acoustic features, such as pitch, tone, and rhythm, can be highly expressive, but children's voices often exhibit greater variability due to developmental differences [12], making emotion more complex. Similarly, visual features—such as facial expressions and movements—play a critical role in conveying emotion but may vary significantly depending on age, expressiveness, and recording conditions. To address these challenges, we propose a multimodal network that integrates multiple modalities, including visual features, acoustic features, and textual data. Additionally, we employ a reinforcement learning-based feature selection approach for acoustic features.

II. Methodology

A. Feature Extraction

Emotion recognition involves identifying, perceiving, and interpreting human behavior and intentions using various modalities, such as visual signals, acoustic properties, and textual information. Similar to humans naturally integrating multiple sources of information to infer emotions, our study incorporates acoustic features, mel-spectrograms, and text modalities.

- *Acoustic features:* Pitch, loudness, frequency, and MFCCs are essential for identifying emotions in children [13]. To extract these features, we employ OpenSmile, an open-source toolkit for audio processing, along with the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [14].

- *Visual features*: The face discloses valuable information during one's emotional responses, making facial expressions a vital source for emotion recognition. In our study, we utilized facial landmark detection and Action Unit (AU) analysis, which capture subtle muscle movements and facial expressions. Additionally, head pose estimation is used to infer the orientation of the face.

- *Textual features:* To integrate text as one of the modalities in our model, we transcribed the audio files using the Whisper model, an advanced automatic speech recognition (ASR) system. Whisper is a multilingual ASR system trained on a diverse range of multilingual and multitask supervised data from the web. It utilizes an end-to-end encoder-decoder Transformer architecture, allowing it to transcribe speech in various languages and even translate those languages into English.

- *Spectrogram features:* For each audio file, a set of Mel-spectrogram representations are generated. The audio is initially processed using the Short-Time Fourier Transform (STFT) with a window length of 500 milliseconds and a hop size of 500 milliseconds. Librosa library was used for generation.

B. Acoustic Feature Selection

We applied Deep Q-Network (DQN), a reinforcement learning technique that dynamically selects features by optimizing a policy to maximize the classification performance. The agent selects a subset of acoustic features, then the CNN model is trained on this subset, and its classification performance (accuracy) is used as the reward signal. This reward is fed back to the agent, allowing it to assess the effectiveness of its selected features. Over time, the agent learns which feature subsets lead to the highest classification performance, ultimately improving the feature selection process.

From the total of 88 eGeMAPS extracted parameters, we utilized 50 selected optimal features emphasize parameters that capture variability in speech such as loudness parameters (loudness_sma3 _amean, meanFallingSlope), pitch parameters, spectral (spectralFlux_sma3 _amean), and MFCCs parameters (m f cc1 − 3_sma3_amean).

C. *Proposed Model*

Figure 1 illustrates the architecture of the proposed multimodal network. The model consists of three distinct networks, each dedicated to a specific modality: visual, acoustic, and text. The features from these three modalities are then fused (concatenated) to form a unified feature vector through the late fusion method. Finally, the model predicts the sample labels through fully connected layers.

From the visual modality, the features are structured as fixed-length vectors, making them well-suited for processing with a Multilayer Perceptron (MLP). MLPs, composed of stacked dense layers, are effective at modelling nonlinear relationships in structured data. By applying fully connected layers, the model captures complex combinations of facial and positional cues that are indicative of emotional states. For the audio modality, we fed acoustic features to the dense layers that learn patterns relevant to emotion expression in speech. MLPs are particularly suitable in these contexts because the features are aggregated across time frames and lack strong spatial or temporal locality. For the case of text modality, we employ a pre-trained BERT model to extract text feature vectors. The text is tokenized, with each token transformed into a vector representation that incorporates positional information to capture both the meaning and its position within the sequence. Then, the token vectors are passed through the self-attention mechanism of the encoder, which evaluates the relevance of each word to others, enabling the capture of contextual dependencies and interactions across the entire text sequence. Representations from unimodal networks are concatenated as follows:

$$\tilde{M}_{v,a,t} = \tilde{U}_v \oplus \tilde{U}_a \oplus \tilde{U}_t$$

Finally, the emotion predictions $\hat{Y}$ are obtained after applying dropout to the joint representation, followed by a fully connected layer, then passing through a Sigmoid activation function.

III. Experiments And Results

A. *Dataset*

We utilized the EmoReact dataset [15], which comprises YouTube videos of children responding to various topics, including food and technology. Each video is divided into clips ranging from 3 to 21 seconds in length, with each clip focusing on a single child's reaction. The children in the videos engage in five activities: being shown the context, answering a question about it, responding to a question, being told a fact, and expressing their opinions. The children are aged 4 to 14 years from diverse racial backgrounds and both genders. In total, the dataset includes 1,102 video clips, pre-split into three sets: 432 clips for training, 303 for validation, and 367 for testing. In our experiments, we employed a 5-fold cross-validation method. Table 1 shows the data statistics.

TABLE I
EmoReact Data Statistics

Emotions	Training	Validation	Testing
Curiosity	148	106	131
Uncertainty	132	97	115
Excitement	144	93	118
Happiness	224	169	211
Suprise	114	79	105

B. *Experimental Results*

The experimental results in Table 2 highlight the influence of different modality settings and feature selection on the weighted F1-score (wF1). Specifically, applying DQN-based feature selection led to an accuracy of 57.84%, marking a significant 2.02% improvement over the baseline model without selection. This result underscores the effectiveness of DQN in optimizing feature representations by reducing dimensionality and filtering out noisy or redundant information. By refining the feature space, DQN enables the model to focus on more discriminative acoustic patterns, ultimately enhancing its ability to recognize emotional states more accurately.

The unimodal models show varying performance levels, with the SVM model trained on acoustic features achieving a higher F1-score of 57.84%, outperforming the CNN model trained on mel-spectrogram. This suggests that acoustic features capture more discriminative emotion-related information [11,14] than mel-spectrogram representations alone in this task.

When introducing text features alongside Mel-spectrogram features, performance improves significantly, with an F1-score of 56.44%. Text enhances emotion detection by providing explicit semantic cues that mel-spectrograms alone cannot capture. While spectrograms convey prosody, pitch, and rhythm, text directly expresses emotions through words, helping to disambiguate similar-sounding emotions. For example, some emotions (e.g., uncertainty versus curiosity) exhibit similar acoustic patterns but can be differentiated based on textual content. In the utterance, "*I guess that might be right…*", a text-based model can understand that "guess" and "might" suggest uncertain emotion.

TABLE 2
Experimental results on different modality settings, where A, V, I, and T represent acoustic, visual, mel-spectrogram, and text features, respectively. The * denotes the result obtained without the Feature Selection.

Model	Modality	wF1 (%)
RBF SVM	A	55.82*
		57.84
MLP	V	59.96
CNN	I	53.98
CNN-BERT	I+T	56.44
Proposed model	V+A+T	**66.15**

Finally, integrating the three modalities improves performance to 66.15%, demonstrating the benefit of incorporating multiple data sources. When comparing the performance of the acoustic modality with multimodal combinations (V+A+T), results highlight the significant impact of acoustic and visual features in emotion detection. They play a crucial role in capturing essential emotional elements. However, combining multiple modalities leads to more accurate detection, as this approach leverages the unique strengths of each modality in understanding and identifying emotions.

Fig. 2 highlights the varying effectiveness of different modality combinations in predicting emotions. Comparing the unimodal models, the MLP model using visual performs well in detecting Curiosity (73.79%) and Surprise (62.86%) compared to spectrogram or acoustic, but struggles with Uncertainty (46.16%) and Excitement (54.13%). The SVM model trained on acoustic features improves performance in Uncertainty (48.10%) and Excitement (58.24%), suggesting that acoustic signatures of these emotions are well expressed through subtle changes in timing, rhythm, or other prosodic features that are not prominently displayed in visual features.

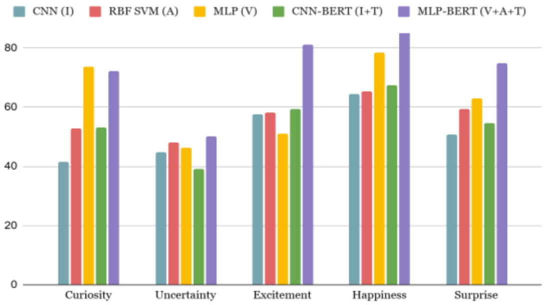


Fig. 2 Modality performance in predicting emotions

The proposed model further improves Excitement and Uncertainty prediction, indicating that the integration of visual, acoustic, and text provides a more comprehensive emotional representation. Interestingly, happiness remains relatively stable across all models, with slight improvements in the multimodal settings, suggesting that this emotion is more easily captured across different feature modalities. This consistency may be attributed not only to the larger number of samples but also to the fact that happiness presents more universal and easily detectable cues across visual, textual (e.g., expressions like 'wow,' 'ha-ha,' 'yeah'), and auditory domains (F0, MFCCs, formant frequencies, etc.). The low accuracy in detecting Uncertainty may be due to ambiguous signals such as hesitations, pauses, or mild vocal modulations that are harder to distinguish.

IV. Conclusions

This study addressed the challenge of emotion recognition in children's speech by proposing a multimodal deep learning framework that integrates audio, visual, and textual data. After applying feature selection in the acoustic modality, a 2.02% performance increase was observed with the DQN model, demonstrating its effectiveness in refining feature representations by reducing redundancy and filtering out noise. Furthermore, the proposed MLP-BERT architecture achieved an F1-score of 66.15% when combining all three modalities, outperforming single-modality approaches. This improvement underscores the advantages of multimodal fusion in capturing subtle emotional cues and improving classification accuracy.

To further advance this research, we plan to extend our study by incorporating additional modalities such as video or facial expressions. Given the significant influence of architecture settings and fusion techniques observed in our experimental results, we will further explore alternative configurations to optimize performance. We aim to extend this approach to real-world applications, including educational tools and mental health monitoring systems, to enhance emotional awareness and intervention strategies.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176)

References

[1] P. Song and W. Zheng, "Feature selection-based transfer subspace learning for speech emotion recognition," IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 373–382, 2018.

[2] P. Ekman, "An argument for basic emotions," Cognition & emotion, vol. 6, no. 3-4, pp. 169–200, 1992.

[3] R. Plutchik, "A general psychoevolutionary theory of emotion," in Theories of emotion. Elsevier, 1980, pp. 3–33.

[4] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "Mexican Emotional Speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody," Data, vol. 6, no. 12, p. 130, Dec. 2021.

[5] A. Mehrabian and S. R. Ferris, "Inference of attitudes from nonverbal communication in two channels." Journal of consulting psychology, vol. 31, no. 3, p. 248-252, 1967.

[6] A. Matveev, Y. Matveev, O. Frolova, A. Nikolaev, and E. Lyakso, "A neural network architecture for children's Audio–Visual emotion recognition," Mathematics, vol. 11, no. 22, p. 4573, 2023.

[7] H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9.

[8] S. Anand, N. K. Devulapally, S. D. Bhattacharjee, and J. Yuan, "Multi-label emotion analysis in conversation via multimodal knowledge distillation," 2023, pp. 6090–6100.

[9] Q. Li, Y. Liu, Q. Liu, Q. Zhang, F. Yan, Y. Ma, and X. Zhang, "Multi-dimensional feature in emotion recognition based on multi-channel EEG signals," Entropy, vol. 24, no. 12, p. 1830, 2022.

[10] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-Visual emotion recognition," Pattern Recognition Letters, vol. 146, pp. 1–7, 2021.

[11] Y. Matveev, A. Matveev, O. Frolova, E. Lyakso, and N. Ruban, "Automatic speech emotion recognition of younger school age children," Mathematics, vol. 10, no. 14, p. 2373, 2022.

[12] L. Neves, M. Martins, A. I. Correia, S. L. Castro, and C. F. Lima, "Associations between vocal emotion recognition and socio-emotional adjustment in children," Royal Society Open Science, vol. 8, no. 11, p.211412, 2021.

[13] G. Cao, Y. Tang, J. Sheng, and W. Cao, "Emotion Recognition from Children Speech Signals Using Attention Based Time Series Deep Learning," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 1296–1300.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr´e, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE transactions on affective computing, vol. 7, no. 2, pp. 190–202, 2015.

[15] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: a multimodal approach and dataset for recognizing emotional responses in children," in Proceedings of the 18th ACM international conference on multimodal interaction, 2016, pp. 137–144.

Recognition of Parent-Child Interaction in Ceiling CCTV Images Using ResNet with UNet

Neunggyu Han[1*] and Yunyoung Nam[2]
[1]*Department of ICT Convergence, Soonchunhyang University, Asan 31538, Korea*
[2]*Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea*
*Contact: az0422sch@sch.ac.kr

Abstract— **In Korea, the importance of diagnosis and treatment of childhood autism spectrum disorder (ASD) has been emphasized recently. Accurate diagnosis and early treatment are most important for ASD. However, existing diagnostic methods can lead to inaccurate diagnostic results due to the intervention of parental judgment. Therefore, it is important to diagnose using more objective indicators. In general, children with autism do not interact well with their parents. Therefore, the presence and rate of interaction with parents can be used as objective indicators. In this paper, we introduce a new method to recognize and classify the presence of interaction between parents and children using deep learning. First, parents and children behave as usual in a place prepared for the experiment. This place is equipped with various toys, and a total of eight CCTVs are installed on the ceiling and floor. We implemented a system to check the presence of interaction using one channel of the data collected in this place. This system uses a different method from existing methods. Existing methods use a method of extracting features using multiple models and then conducting analysis. However, in this paper, UNet and ResNet are used to automatically extract features from images and analyze the presence of interaction. Using this structure, we achieved an accuracy of about 85%. There are some cases where non-interactions are confused for interactions, so improvements are needed in this area. In addition, after the improvements are completed, we will implement a system without blind spots by using all eight channels.**

I. Introduction

Recently, the importance of diagnosis and treatment of childhood autism spectrum disorder (ASD) has been emphasized in Korea. Accurate diagnosis and early treatment are most important for ASD. However, existing diagnostic methods can lead to inaccurate diagnostic results due to the intervention of parental judgment. For example, even if symptoms suspected of ASD are found, they may simply be thought to be trivial due to personality, etc. Therefore, it is most important to diagnose based on objective indicators rather than subjective perspectives. If this diagnostic method is performed only at hospitals, it may not lead to an actual diagnosis due to resistance, etc. Therefore, a system is needed to assist diagnosis at home and inform parents of this. In general, children with ASD do not interact well with their parents. Therefore, the presence or absence of interaction and the rate of interaction can be used to confirm whether ASD is suspected.

There are several ways to determine whether or not ASD is suspected through interaction. Among them, the method of utilizing CCTV does not cause much discomfort to parents and children, so they can observe their usual appearance. Among the methods of processing such images, there is a method utilizing deep learning [1-2]. These existing methods extract features using various models and proceed with analysis based on them. There are several problems when using these methods. First, an object recognition model is used for feature extraction. This model uses a method of recognizing or not recognizing objects based on a certain threshold value. Therefore, feature extraction is inaccurate when an object is in a complex pose or is difficult to recognize. The second is a pose and gaze estimation model that is used as an auxiliary. These models are generally models created assuming that all parts of the body are visible. However, in some compositions, not all parts of the body are visible, so feature extraction fails. To compensate for these problems, multi-channel CCTV is used, but there is a disadvantage that implementation is difficult.

In this paper, we introduce a new method to determine whether there is interaction. This method distinguishes regions using an image segmentation model, rather than an object recognition algorithm that uses thresholds. UNet [3] is used to distinguish regions for input images, and this is used as an attention map and applied to the input image. Then, this image is input to ResNet [4] to analyze whether there is interaction by considering changes over time. Using this method, the phenomenon of recognition being cut off by thresholds is eliminated, and more accurate analysis results can be obtained.

A separate space was prepared for the experiment in this paper. Various toys are prepared in this space. Therefore, natural interaction with parents is possible. In addition, four CCTVs are installed on the ceiling and four CCTVs are installed on the floor. These CCTVs observe parents and children without any blind spots. The dataset used in this paper is not a public dataset but a dataset collected directly. The experiment was conducted on one channel out of the eight collected channels of data. The experimental results classified whether there was interaction with an accuracy of 85%.

II. System Implementation

A. Datasets

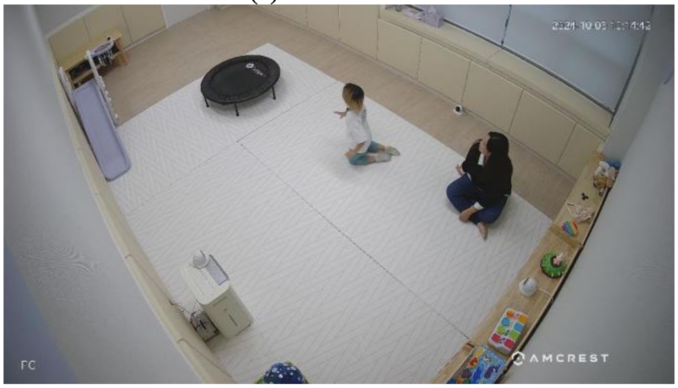
We collected the dataset directly in the experimental space. The collected dataset consists of a total of 8 channels. Of these, 4 channels are ceiling CCTVs and the remaining 4 are floor CCTVs. The CCTV footage has a frame rate of 15 frames per second and a resolution of 3840x2160. In addition, one parent and one child appeared in each video to conduct the activity. With this, 16 video data per channel, a total of 128 video data were collected. Among the collected video data, the experiment was conducted using the video data for channel 1. The video recorded in this channel can be seen in Fig. 1. This data is the optimal channel for checking all the toy facilities from the entrance door of the experimental space. Of these video data, 11 were used for learning and 5 were used for verification.

Preprocessing of video data was performed as follows. First, frame data was extracted from the video at 3 frames per second. This was done considering the capacity and the amount of change between each frame. If the amount of change between frames is small, it is difficult to analyze the actual movement.

Next, an example of labeling is shown in Fig. 1. Each frame data was directly checked, and if the parent and child were doing something together in close proximity, it was labeled as interaction. Conversely, if the parent and child were doing something separate from each other, it was labeled as non-interaction. Finally, if there was no object to be recognized, it was labeled as none.

The ratio of labeled frame data is as follows. The total frame data of training data is 22,103, with 12,209 interactions, 8,314 non-interactions, and 1,580 none. The total frame data of validation data is 9,861, with 6,187 interactions, 3,323 non-interactions, and 351 none.

(a) Interaction



(b) Non-interaction

(c) None

Fig. 1 Those images show example of interaction, non-interaction, and none situations in CCTV channel 1.

B. UNet and ResNet based Interaction Recognition Model

Other studies use methods that extract features using multiple models and analyze them. However, such methods may lose some of the features of the original image. Therefore, this paper does not use a separate model to extract features, but only uses UNet and ResNet to recognize whether there is interaction.

In this paper, we recognize the interaction by constructing a model as shown in Fig. 2. First, UNet and 2D-ResNet are used to extract vectors of input images. UNet has a simple structure and consists of 18 layers in total. And it outputs 4 attention maps. This attention map is applied to the input image as bilinear attention and input to 2D-ResNet. And 2D-ResNet consists of 54 layers and a bottleneck is applied to the residual block. In addition, it consists of 1 convolutional layer, Global Average Pooling layer, and fully connected layer to convert to vector. The activation of the last fully connected layer is hyperbolic tangent. Finally, the vectors are input to 1D-ResNet. 1D-ResNet consists of 27 layers in total and down sampling is performed twice. And finally, the interaction is classified.

Applying attention to images with UNet, like this model, allows the model to learn by reducing information about unnecessary parts and selecting only necessary information. In other words, higher performance can be achieved compared to when images are input without being processed.

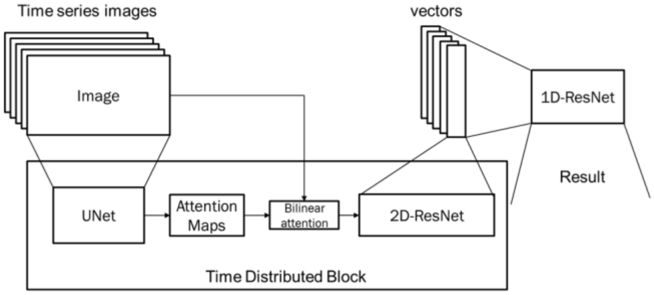


Fig. 2 Model Architecture

C. Training for Model

Model learning was performed using parameters such as TABLE. 1. The input size is a value determined by considering the GPU memory usage and processing speed. A value larger than the input size of 224x224 can also be used. However, the

GPU memory usage increases and the learning speed slows down. The window size is an optimal value obtained through experiments. If this value exceeds 32, it is difficult to use due to GPU memory usage issues and the performance decreases. In addition, if it is set to less than 16, the model performance deteriorates significantly.

For data augmentation, vertical translation, left-right inversion, noise, and brightness adjustment were applied. Among these, vertical translation and left inversion are applied on a window basis. Noise and brightness adjustment are applied to each image. However, noise and brightness adjustment are set so that there is not a large difference compared to the original. Finally, the data input to learning is not sequentially, but is continuously taken from a random location as much as the window size and then used.

TABLE. 1 Training Parameters

Parameter	Value
Input size	224x224
Batch size	16
Learning rate	0.001
Optimizer	Adam
Window size	16

III. Experimental Result

A. Experimental Setup

The experimental environment in which model learning and performance evaluation were conducted in this paper was as shown in TABLE. 2.

TABLE. 2 Training and Evaluate Environment

Deep Learning Server	
CPU	Xeon Silver 4216 x2
Memory	192GB
GPU	RTX A5000 x3
OS	Ubuntu 22.04

B. Evaluation of model

After training with the dataset of this paper, the model is evaluated. The accuracy of the model is 85%. Also, the confusion matrix of the model is as shown in Fig. 3. The order of the classes is interaction (0), non-interaction (1), and none (2). Among these, there are some cases where the non-interaction class is incorrectly recognized as interaction. The cause of this confusion seems to be that the labeling is somewhat unclear. To solve this problem, we may need to consider other indirect classes other than interaction and non-interaction. In other words, we classify a certain action and then distinguish whether it is interaction or not based on this. In addition, there is a problem of data loss caused by the small input size. The input size of 224x224 is small, so data such as face or gaze cannot be confirmed. It is necessary to reduce the usage of GPU memory, which is the cause of this problem, and increase the input size.

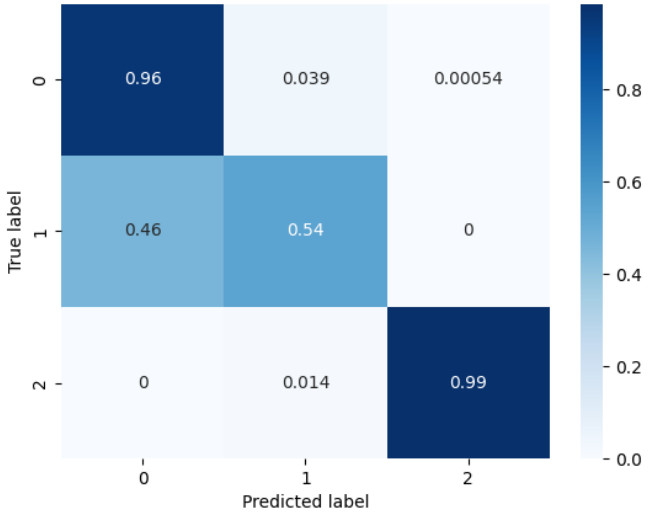


Fig. 3 Confusion matrix

C. Performance Comparison

In this paper, the experiment was also conducted. Among them, the highest performance in the method of using 2D-ResNet with UNet and 1D-ResNet was the highest performance.

As we experimented, we often found overfitting phenomena. Therefore, a simple model was organized to configure a model with high performance. But the simpler model was, the less performance of the model was. Therefore, in this problem, it was important to use the appropriate model rather than using a simple model.

TABLE. 3 Performance Comparison

Method	Accuracy
2D-ResNet18 with UNet and 1D-CNN	66.5%
Simple 2D-CNN with 1D-CNN	66.6%
3D-ResNet18 Only	71.5%
3D-ResNet18 with UNet	82.5%
2D-ResNet with UNet and 1D-ResNet	85.0%

D. Verification of the Attention-Applied Image and Discussion

By analyzing the images to which attention is applied through UNet, we can see which parts the model considers important. First, Fig. 4 shows the images to which attention is applied. Since four attention maps are generated, they are directly applied to four images. In addition, the images used in the dataset are images used for verification. And these images are interactive images.

In Fig. 4, the parts with low importance are marked in black, and the parts that are not are marked with the original data. This image shows which parts the model considers important. First, parts such as walls are marked in black because parents and children do not approach them. As for the middle area, you can see that none of the channels are painted in black except for one. The one channel painted in black seems to indicate where non-interactions generally occurred, and the rest seem to be about interactions. Finally, in the two channels, you can see that the floor area is wrapped in black, but it is not painted in black based on the parent. This shows that the model determines whether there is an interaction based on the parent. This shows that the model can achieve high performance because it directly determines specific conditions and appropriately divides the area.

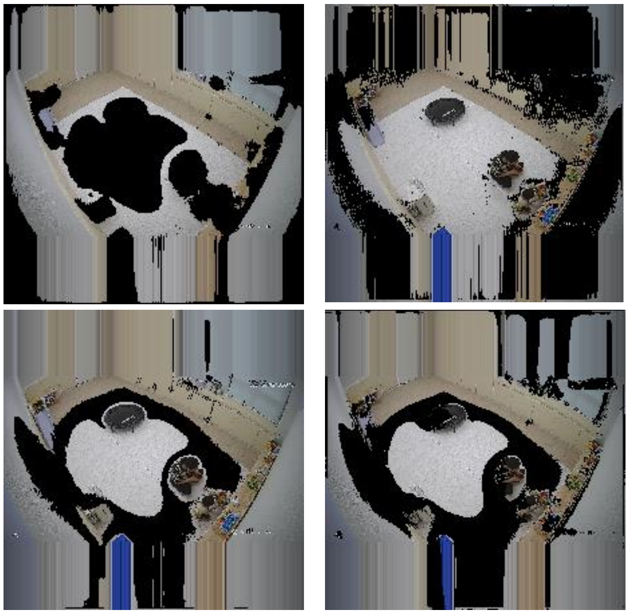


Fig. 4 Those images are shows attention-applied images

IV. Conclusions

The current ASD diagnosis method relies on the opinions of parents, which has the problem of low diagnostic accuracy. Therefore, it is very important to create an objective index. Among the objective indexes, there is the interaction between parents and children. In this paper, a new deep learning-based model that analyzes the interaction using CCTV images in an experimental space was constructed. This model is constructed using UNet and ResNet. In addition, an accuracy of 85% was achieved based on the dataset that was constructed directly.

There are several problems at this stage. First, we were able to confirm that non-interactions are confused with interactions. The cause of this confusion is the ambiguity between interactions and non-interactions. Most of the data labeled as interactions are in close proximity to the parents and children. However, there are also cases where non-interactions appear even when they are close to each other. In such situations, we were able to confirm that interactions and non-interactions are confused. Therefore, we should consider a method to distinguish these two as an indirect class other than simply interactions and non-interactions. In addition, the structure of the model should be optimized so that it can use a larger input size.

Finally, at the current stage, the model was constructed and the experiment was conducted using only one channel of video. However, a total of eight cameras were installed in the experimental space. Therefore, the plan is to implement it so that analysis can be performed from all directions without any blind spots by using all cameras.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176)

References

[1] Nikbakhtbideh, B. (2023). An AI-based Framework For Parent-child Interaction Analysis.

[2] M. Cheng et al., "Computer-Aided Autism Spectrum Disorder Diagnosis With Behavior Signal Processing," in IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 2982-3000, 1 Oct.-Dec. 2023, doi: 10.1109/TAFFC.2023.3238712.

[3] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234-241). Springer international publishing.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[5] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Text-Guided Generation of Child-Friendly 3D Objects for Virtual Environments

G. Kim[1], Y. Jung[2], W. Shin[2], and S. Seo[1,*]
[1]*School of Art and Technology, Chung-Ang University, Anseong-si, 17546, South Korea*
[2]*Department of Applied Art and Technology, Chung-Ang University, Anseong-si 17546, South Korea*
*Contact: sanghyun@cau.ac.kr, phone +82-10 7273 0318

Abstract—**The design of modern VR content lacks consideration for the emotional and cognitive characteristics of young children, in addition to accessibility for non-experts. In this paper, we establish a pipeline that enables the generation and application of child-friendly VR objects using a prompt-based AI system without requiring expertise in 3D modeling. To achieve this, text prompts incorporating Norman's visceral design principles were utilized to generate 2D images that reflect child-friendly elements such as color and shape. These images were then converted into 3D models using Microsoft TRELLIS, a Structured Latent Representation (SLAT)-based system. The resulting 3D objects were designed for integration into VR content for early childhood education. Future research will focus on evaluating the effectiveness of these generated objects in reflecting child-friendly elements and optimizing their usability in real-world educational environments. This paper is expected to serve as a foundational resource for the development of child-oriented content.**

I. Introduction

In recent years, the application of Virtual Reality (VR) in educational programs and play-based content for young children has been actively expanding. VR-based educational content is recognized as an effective tool for promoting sensory development and creative thinking in young children by providing a highly immersive and intuitive learning experience [1]. However, a significant portion of currently developed VR educational content relies on general 3D objects without fully considering the cognitive and emotional characteristics of different age groups. This design approach fails to adequately reflect children's visual preferences and emotional responses, potentially leading to reduced engagement and limited learning effectiveness. Furthermore, conventional methods often require specialized 3D modeling skills, posing a significant barrier for general educational content developers. Therefore, there is a growing need for an automated solution that enables non-experts to easily generate and integrate child-friendly virtual objects into VR environments.

This study proposes a pipeline utilizing GPT-4-based prompt-driven image generation technology to automatically create and implement child-friendly 3D objects. The proposed pipeline allows educational content developers to generate 3D objects that incorporate child-friendly design elements using only text input and seamlessly place them within virtual environments. Through this approach, developers of VR educational content for young children can effectively construct immersive learning environments without prior knowledge of 3D modeling and design. This is expected to enhance both the accessibility and usability of virtual educational content tailored for early childhood learning.

II. Methodology

This study proposes a virtual object generation method for developing child-friendly VR content by utilizing Norman's Affective Design theory, particularly the concept of Visceral Design, to evoke positive emotions and experiences in young users [2]. To achieve this, a text prompt-based pipeline was developed for generating 2D images and converting them into 3D models. The structure of the proposed pipeline is summarized in Fig. 1 and consists of three main stages: (1) prompt input with child-friendly keywords, (2) generation of 2D images, and (3) 3D model generation based on the 2D images.

A. *Selection of Child-Friendly Design Criteria*

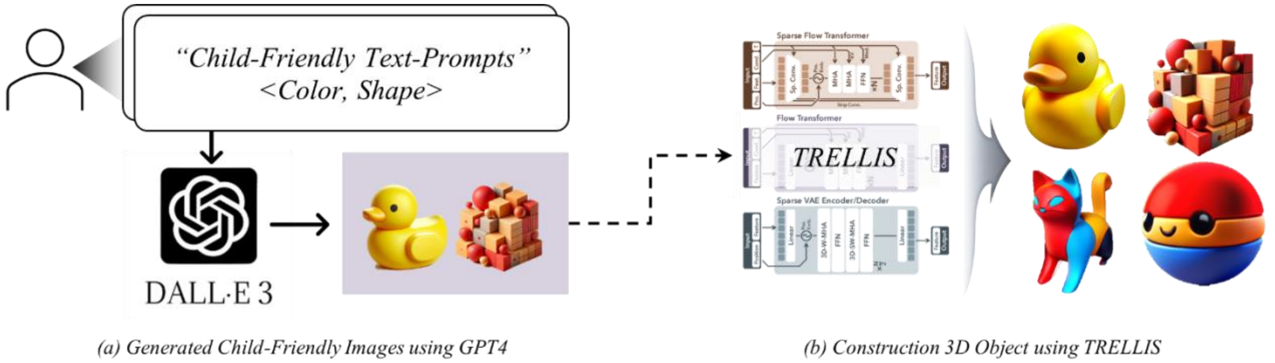
To ensure that the 3D models generated through prompt-based methods are intuitively familiar and engaging for young children when integrated into actual content, design criteria based on children's shape and color preferences were established. According to relevant studies, young children tend to prefer rounded shapes such as circles and spheres, as they resemble familiar objects in their surroundings, including cars, a mother's face, and the sun. Additionally, children favor simplified forms over realistic representations and exhibit a strong affinity for animal-like shapes, particularly those of cats, rabbits, dogs, lions, butterflies, and bears.

In terms of color preference, children are generally attracted to warm tones, neutral tones, and highly saturated pure colors, while they tend to avoid dark and muted shades. Furthermore, as young children begin to develop color concepts by recognizing the primary colors of frequently encountered objects, the use of primary colors may be necessary to accommodate those who have not yet fully formed color perception [3].

TABLE I
Child-Friendly Color and Shape Design Criteria

Child-Friendly Design Criteria	
Color	**Shape**
Primary Colors	Round Shapes and Curves
Warm Colors (Yellow, Orange, Red)	Simple Shapes (e.g., Circle, Triangle)
Avoidance of dark and muted colors	Animals-like Shapes

B. *Integration of TRELLIS 3D for Enhanced Object Generation*



(a) Generated Child-Friendly Images using GPT4

(b) Construction 3D Object using TRELLIS

Fig. 1 Child-Friendly 3D Object Generation Pipeline: (a) AI-Generated Images using GPT-4, (b) 3D Object Construction via TRELLIS[4]

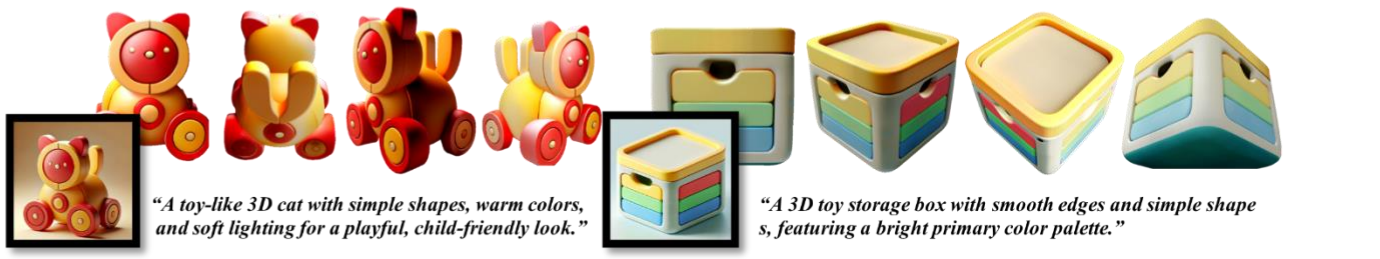

"A toy-like 3D cat with simple shapes, warm colors, and soft lighting for a playful, child-friendly look."
"A 3D toy storage box with smooth edges and simple shape s, featuring a bright primary color palette."

Fig. 2 Child-friendly 3D assets only built based on text prompts

TRELLIS 3D employs Structured Latent Representation (SLAT), enabling high-quality 3D asset generation with efficient computational performance. This model allows for:

- **Versatile 3D Asset Generation:** Supporting multiple output formats such as meshes, Gaussians, and Radiance Fields.
- **Scalable and Flexible Editing:** Enabling fine-tuned structural modifications and local 3D adjustments.
- **Efficient 2D-to-3D Conversions:** Optimizing child-friendly design principles through structured latent encoding.

By integrating TRELLIS 3D into our pipeline, we enhance the ability to generate customized 3D models that maintain both geometric accuracy and aesthetic appeal [4].

C. *Text Prompt-Based 2D Image Generation*

To generate child-friendly 2D images, the GPT-4-based DALL·E 3 model was utilized. Prompts were crafted to include keywords reflecting shapes and colors preferred by young children, and 2D images were generated using GPT-4 based on these prompts [5].

D. *2D-to-3D Model Conversion*

The generated 2D images were converted into 3D models using Microsoft TRELLIS. TRELLIS is a tool that automatically generates 3D models from image data, enabling the rapid creation of high-precision 3D object models. The converted 3D models were optimized for immediate use in VR content development without requiring additional texturing processes [4].

III. Experiment

This study evaluates the applicability of 2D image-based 3D models in a virtual reality (VR) environment by constructing a virtual playroom using Unreal Engine. The main steps of the experiment are as follows:

A. *3D Model Conversion and Environment Setup*

2D images generated using GPT-4-based DALL·E 3 were converted into 3D models using Microsoft TRELLIS. The generated 3D models were placed in a virtual playroom environment within Unreal Engine, as shown in Fig. 3, to create an interactive VR space.

B. *Experimental Environment Setup*

The virtual playroom included the following 3D objects:

- **Furniture:** Bookshelf, Desk, Chair, Board
- **Play equipment:** Slide
- **Toys:** Toy car, Cat figurine, Cube blocks
- **Stationery:** Crayons

Each object was evaluated by comparing it to its original 2D conceptualization to determine how accurately it was recreated in the 3D space. The objects were placed to allow children to interact naturally within the virtual playroom. Fig. 3 illustrates the layout and placement of these objects within the Unreal Engine environment.

C. *Experimental Results and Analysis*

- **Model Reproducibility:** The expected 3D forms from the 2D images were accurately reflected in the 3D environment, demonstrating immediate applicability.
- **Object Intuitiveness:** The colors and shapes designed to be familiar to children were well preserved, resulting in high emotional satisfaction based on Visceral Design principles.
- **Practical Usability:** The objects were easily adjustable in size and placement within the VR environment, requiring minimal additional texture modifications for application.

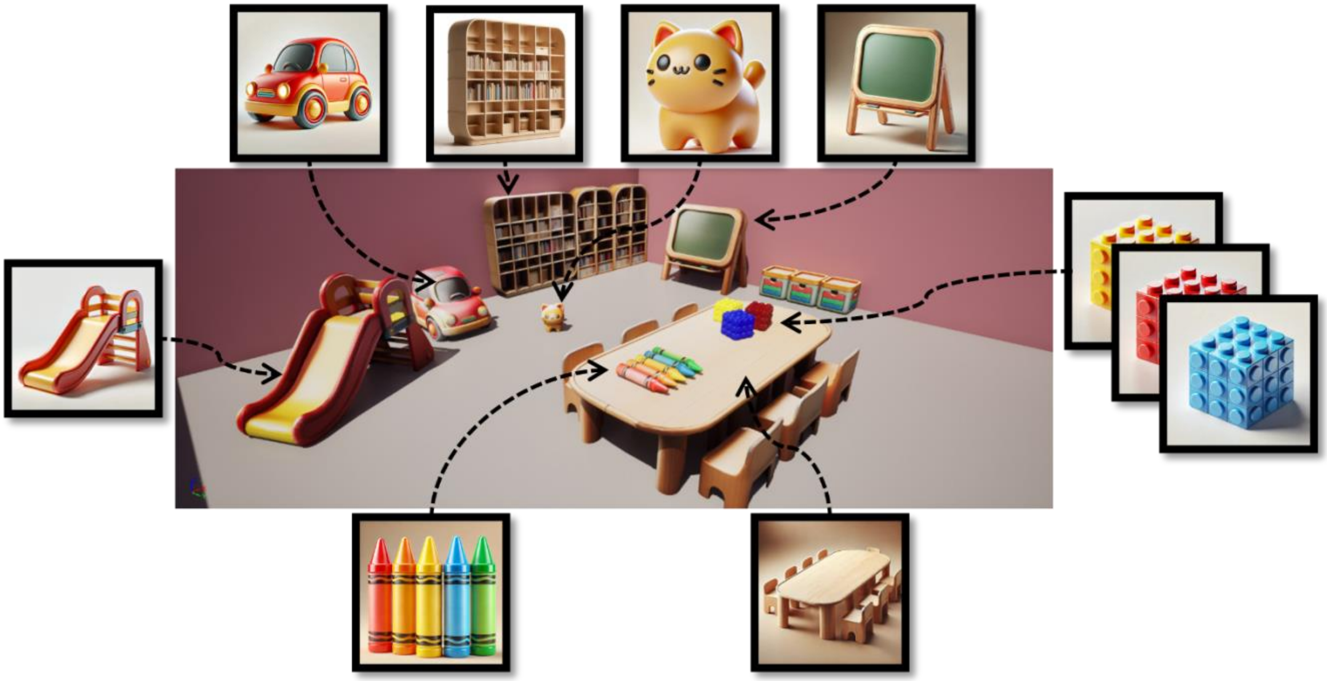


Fig. 3 Applying 3D models converted from 2D images to game engine (Unreal Engine 5)

Through this experiment, we confirmed that 2D image-based 3D objects can be effectively applied in a VR environment. Future research will focus on evaluating real user responses and analyzing children's interaction data to optimize design usability.

IV. Results

The 3D objects generated in this study were designed with child-friendly shapes and colors. Utilizing the prompt-based 3D model generation pipeline illustrated in Fig. 1, 3D models tailored to specific age characteristics could be created with simple text input. Fig. 2 presents the results produced using the proposed method, demonstrating that this approach is more user-friendly and yields faster outcomes compared to 3D modeling techniques (e.g., Blender, Maya).

Unlike conventional 3D modeling methods, which require mastering complex software and performing manual tasks, the proposed approach enables the immediate generation of ready-to-use 3D models solely through prompt inputs, highlighting its distinct advantages. Furthermore, as shown in Fig. 2, 3, the generated 3D models exhibit forms similar to objects commonly used in real-world early childhood educational environments. This suggests that generative prompt-based 3D object creation technology can effectively reflect the preferences of specific age groups, making it a valuable tool for child-friendly VR content development.

V. Conclusions

This paper proposed a prompt-based automated pipeline for generating child-friendly 3D object models, designed for use in virtual environments. By employing this approach, non-experts without prior 3D modeling skills can easily participate in VR content development. Specifically, by establishing child-friendly design criteria and applying relevant keywords to text prompts, the proposed pipeline enables the automatic creation of 3D object models that are both familiar and educationally beneficial for young children.

Future research will focus on evaluating the impact of the generated objects on actual child users. Eye-tracking data analysis will be used to measure visual engagement, while surveys will assess emotional responses. Additionally, user feedback will be collected to refine prompt design. Further studies will also explore segmenting young children into specific age groups to analyze their distinct characteristics and preferences. Based on these findings, a tailored model generation approach will be developed, considering both age-specific traits and gender-based design preferences.

To achieve this, future research will incorporate children's preferences into the training data and conduct experiments to validate the generated models within real virtual environments.

This paper is expected to serve as foundational research for child-centered virtual content development. As generative AI-based 3D model generation technologies continue to advance, they are anticipated to be creatively and effectively utilized across various educational and play environments.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176, Contribution Rate: 50%) and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2023 (Project Name: Development of digital abusing detection and management technology for a safe Metaverse service, Project Number: RS-2023-00227686, Contribution Rate: 50%)

References

[1] F. Zhang, Y. Zhang, G. Li, and H. Luo, "Using Virtual Reality Interventions to Promote Social and Emotional Learning for Children

and Adolescents: A Systematic Review and Meta-Analysis," Children, vol. 11, no. 1, p. 41, 2023. DOI: 10.3390/children11010041.

[2] D. Norman, *Emotional Design: Why We Love (or Hate) Everyday Things*. New York, NY, USA: Basic Books, 2004.

[3] J. Y. Jung, J. I. Han, and S. G. Woo, "A study on textile design development for children's furniture," *Journal of Korean Design Culture Society*, vol. 19, no. 4, pp. 630-640, Dec. 2013.

[4] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, "Structured 3D Latents for Scalable and Versatile 3D Generation," arXiv preprint arXiv:2412.01506, 2024

[5] OpenAI, "DALL·E 3: Text-to-image generation with advanced capabilities," OpenAI, 2025. [Online]. Available: https://openai.com/dall-e-3.

Optimization of Machine Learning Algorithms for Crosstalk Analysis

Minhyuk Kim
Department of Electronic Engineering, Soonchunhyang University, Asan, 31538, Korea
Contact: MH.Kim@sch.ac.kr, phone +82-42-530-1326

Abstract— **Processor speeds are increasing rapidly and digital systems require significant data bandwidth as technology advances. This requires careful consideration of signal integrity to ensure reliable, high-speed data processing. Due to the high level of integration, crosstalk has become an important area of signal integrity research for electronic packages. In this study, analytical formulas were analyzed to identify the characteristics that can predict crosstalk in multi-conductor transmission lines. Through the analysis, five variables were found, and a data set consisting of 302,500 data points was obtained. This study evaluated performance of various regression models for automatic machine learning optimization by comparing machine learning predictions with analytical solution. Extra tree regression consistently outperformed other algorithms, with coefficients of determination above 0.9 and root mean square logarithmic errors below 0.35. The study also notes that different algorithms produced different predictions for the two metrics.**

I. Introduction

Signal integrity problems, previously addressed mainly from a radio frequency standpoint, now pose a crucial issue for both digital and analog designs as technology advances and digital systems become faster.

Crosstalk analysis is typically performed during the design phase using physical-based (PB) approaches [2, 3]. There has been extensive research into the analysis of crosstalk using formulas [4, 5]. However, implementing these formulas for complex real-world problems is not straightforward. Additionally, calculating multiple integrals for computing bent shapes is much faster than PB methods, but still requires a considerable time. PB techniques are relatively precise, they are also slow, prompting researchers to explore machine learning (ML) alternatives [6-9]. Researchers employed ML in [6] to analyze signal and power integrity in various experiments, with a significant portion of the training data obtained through HFSS. ML research applied to a variety of electromagnetic problems still relies on traditional PB methods that are sluggish, requiring significant computing resources and time for data acquisition. As a result, progress in this field is relatively slow

An analytical approach was used to conduct crosstalk analysis of transmission lines with multiple conductors to gather data for optimizing machine learning (ML) algorithms. The detailed analytic solutions for these transmission lines is summarized in [4] and has been implemented in several studies [5, 10]. In addition to the theoretical solution for parallel transmission lines discussed in [4], an analytical solution for crosstalk in bent structures has been examined in [5]. Consequently, it is feasible to obtain crosstalk datasets for diverse structures using solely theoretical solutions, which can be acquired rapidly.

In this study, a dataset was constructed by analyzing the analytical formula and extracting the essential parameters for learning. The outcomes were obtained through a ML and compared to the analytical solution using widely used regression analysis metrics. The most optimized ML regression algorithm for the model used in this study was determined using automatic machine learning (AutoML).

II. Analytic Formula of Multiconductor Transmission Lines

Fig. 1 demonstrates that crosstalk can be divided into two categories: near-end crosstalk (NEXT), where the induced current propagates in the opposite direction of the signal current, and far-end crosstalk (FEXT), where the induced current propagates in the same direction.

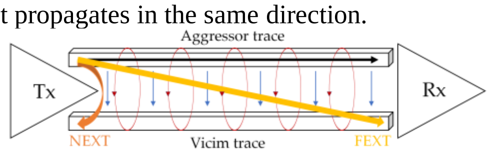


Fig. 1 NEXT/FEXT between parallel traces.

Fig. 2 depicts the model derived from this analysis. The height of the board is represented by t, and its dielectric permittivity is denoted by ε_r. The width and length of the trace are indicated by w and L, respectively. The distance between the traces is denoted by s.

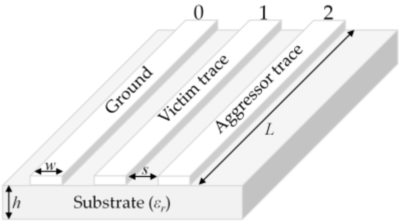


Fig. 2 Crosstalk analysis model.

In order to examine crosstalk using machine learning, the dataset is organized as outlined in Table 1.

TABLE I
Dataset for Analyzing Crosstalk

Parameters	Min.	Max.	step
Frequency [MHz]	100	1000	100
t [mm]	0.5	1.5	0.1
ε_r	4	5	0.1
w [mm]	0.1	0.5	0.1
s [mm]	0.1	0.5	0.1
L [m]	0.1	1	0.1

Of the 302,500 data points, 70% were utilized for training and 30% for evaluating the model.

III. Open-source, low-code machine learning library

The prediction of ANNs is influenced by data preprocessing, feature selection, and the training algorithm used. Popular regression analysis algorithms in this field include Linear Regression, Ridge, and Lasso, with ongoing research efforts dedicated to their effectiveness. In the case of replacing PB methods with ML, the difference in results depending on the training algorithm has been studied in the field of wireless power transmission [10]. However, in signal integrity studies, the number of algorithms used is not many, or it is difficult to identify which algorithm was used.

In this study, AutoML was utilized to efficiently assess the accuracy of numerous machine learning algorithms. Traditional ML approaches entail significant difficulties in constructing ML models, which involve intricate and repetitive tasks, such as data preprocessing, model selection, and hyperparameter tuning. AutoML streamlines these processes by automating them as much as possible, thereby boosting efficiency at each stage of ML and deep learning (DL).

IV. Results and Discussion

The study assessed prediction accuracy and algorithm performance for NEXT and FEXT through the use of root-mean-square log error (RMSLE) and coefficient of determination (R^2) metrics. In regression analysis, the RMSLE is a reliable measure for outlier detection due to its logarithmic scale and emphasis on relative error. Also, R^2 established the correlation degree between the regression line and data points, often utilized for comparative purposes of relative performance.

The RMSLE and R^2 evaluation results for all 19 regression algorithms. The model's performance is better when RMSLE is closer to 0 and R^2 approaches 1. Of all machine learning models, et and rf exhibited superior results on metric (9) among various regression algorithms. The evaluation using et yielded slightly better outcomes, thus the models were assessed using non-training data. The evaluation results are shown in Table 2.

TABLE 2
Model evaluation result using et

Crosstalk	RMSLE	R2
Magnitude (NEXT)	0.0068	0.9770
Angle (NEXT)	0.2672	0.9041
Magnitude (FEXT)	0.0048	0.9824
Angle (FEXT)	0.3135	0.9222

The par model does not accurately reflect the trend line in the data, as evidenced by its negative value in the R^2 evaluation results. Because the optimal algorithm may be incorrect and may vary depending on the analysis environment, it is important to identify the optimal algorithm in advance.

The par model does not accurately reflect the trend line in the data, as evidenced by its negative value in the R^2 evaluation results. Because the optimal algorithm may be incorrect and may vary depending on the analysis environment, it is important to identify the optimal algorithm in advance.

V. Conclusions

Relevant parameters were identified through analysis of pre-existing analytic formulas. A dataset was efficiently generated by utilizing an analytic approach. PyCaret, a powerful machine learning library in Python, was employed to predict crosstalk using this dataset. The results from various learning algorithms were evaluated using commonly-used metrics such as RMSLE and R^2 in regression analysis. The evaluation revealed significant variation in prediction performance among the different algorithms. The feature importance of the variables for the printed circuit board model used in this study was calculated, and the optimal algorithm was determined. For future studies aimed at facilitating their application to real-world issues, this research will be expanded to complex models with bent structures, utilizing the algorithms identified as the most effective in this investigation.

References

[1] J. Fan, X. Ye, J. Kim, B. Archambeault, and Orlandi, "Signal Integrity Design for High-Speed Digital Circuits: Progress and Directions," IEEE Trans. Electromagn. Compat., vol. 2, pp. 392-400, 2010.

[2] M. Schierholz, A. Sánchez-Masís, A. Carmona-Cruz, X. Duan, K. Roy, C. Yang, R. Rimolo-Donadio, and C. Schuster, "SI/PI-Database of PCB-Based Interconnects for Machine Learning Applications," *IEEE Access*, vol. 9, pp. 34423-34432, 2021.

[3] G. Shan, G. Li, Y. Wang, C. Xing, Y. Zheng, and Y. Yang, "Application and Prospect of Artificial Intelligence Methods in Signal Integrity Prediction and Optimization of Microsystems," *Micromachines*, vol. 14, p. 344, 2023.

[4] C. R. Paul, Analysis of Multiconductor Transmission Lines, 2nd ed., John Wiley & Sons, New Jersey, 2007.

[5] S. W. Park, F. Xiao, D. C. Park, and Y. Kami, "Crosstalk Analysis Method for Two Bent Lines on a PCB Using a Circuit Model," *IEICE Trans. Commun.*, vol. 90, pp. 1313-1321, 2007.

[6] M. Swaminathan, H. M. Torun, H. Yu, J. A. Hejase, and W. D. Becker, "Demystifying Machine Learning for Signal and Power Integrity Problems in Packaging," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 10, pp. 1276-1295, 2020.

[7] W. Beyene, "Application of Artificial Neural Networks to Statistical Analysis and Nonlinear Modeling of High-Speed Interconnect Systems," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 26, pp. 166-176, 2006.

[8] H. Kim, C. Sui, K. Cai, B. Sen, and J. Fan, "Fast and Precise High-Speed Channel Modeling and Optimization Technique Based on Machine Learning," *IEEE Trans. Electromagn. Compat.*, vol. 60, pp. 2049-2052, 2017.

[9] T. Lu, J. Sun, K. Wu, and Z. Yang, "High-Speed Channel Modeling with Machine Learning Methods for Signal Integrity Analysis," *IEEE Trans. Electromagn. Compat.*, vol. 60, pp. 1957-1964, 2018.

[10] A. Shoory, M. Rubinstein, C. Romero, N. Mora, and F. Rachidi, "Application of the Cascaded Transmission Line Theory of Paul and McKnight to the Evaluation of NEXT and FEXT in Twisted Wire Pair Bundles," *IEEE Trans. Electromagn. Compat.*, vol. 55, pp. 648-656, 2013.

[11] S. A. A. Mahmud, P. Jayathurathnage, and S. A. Tretyakov, "Machine Learning Assisted Characteristics Prediction for Wireless Power Transfer Systems," *IEEE Access*, vol. 10, pp. 40496-40505, 2022.

Comparison of Developmental Levels Between Children with Developmental Disabilities and Typically Developing Children : A Study on Three-Year-Olds

Yaena Ha[1], Chomyong Kim[1], Yunyoung Nam[2,*]
[1]*Emotional and Intelligent Child Care System Convergence Research Center, Soonchunhyang University, Republic of Korea*
[2]*Department of Computer Science and Engineering,, Soonchunhyang University, Republic of Korea*
*Correspondence: ynam@sch.ac.kr, phone +82-041 530 1151

Abstract— **Advancements in virtual reality (VR) technology offer new possibilities for intervention programs aimed at supporting children with developmental disabilities (DD). VR provides safe and repeated experiences of real-life situations, making it an effective intervention tool. This study compared the developmental levels of three-year-old children with DD and typically developing (TD) children using data from the 2023 National Survey of Children's Health (NSCH) in the United States. A total of 262 children were analyzed, including 131 DD children and 131 TD children. The results indicated that children with DD exhibited significantly lower cognitive, language, social, and emotional developmental levels compared to TD children. Based on the results of this study, when developing VR interventions to enhance the developmental levels of children with DD, content should focus on promoting the specific developmental areas in which delays are observed compared to TD children. In addition, differentiated intervention strategies that reflect individual differences among DD children should be established.**

I. Introduction

Advancements in virtual reality (VR) technology have opened new possibilities for intervention programs aimed at enhancing the developmental levels of children with developmental disabilities (DD) [1]. VR technology allows children to safely and repeatedly experience situations that are difficult to encounter in real-life settings, making it an efficient intervention tool for promoting their development [2], [3]. Early childhood, particularly around the age of three, is a critical period for cognitive, language, social, and emotional development. Identifying developmental differences between children with and without developmental disabilities at this early stage can provide valuable insights for designing effective intervention strategies [4], [5]. This study aimed to compare the developmental levels of three-year-old children with DD and typically developing (TD) children to develop interventions utilizing virtual reality technology for enhancing the developmental levels of children with DD.

II. Method

A. Study Participants

The study participants were children with developmental disabilities who had been diagnosed with autism spectrum disorder (ASD) or intellectual disability (ID). The total number of study participants was 262, including 131 three-year-old children with developmental disabilities (DD) and 131 typically developing (TD) children. Therefore, to analyze the data of DD and TD children, 131 children were randomly sampled from the 33,121 TD children.

B. Study Data

This study utilized big data from the National Survey of Children's Health (NSCH) collected in the United States in 2023. The independent variable was whether the study participants were children with DD and TD. The dependent variable was the developmental level of the child, and variables related to the development of three-year-old children were selected from the NSCH data. These included 12 items for measuring cognitive and language development, and 3 items for assessing social and emotional development. The scale for each item was coded so that if the child was always able to perform the item it was defined as able to perform = 1, and if not it was defined as needs support = 2.

C. Statistical Analysis

Descriptive statistics were conducted to analyze the demographic characteristics of the study participants. Chi-square test was conducted to compare the developmental levels of DD and TD children. SAS 9.4 was used for data cleaning and analysis.

III. Results

A. Demographic characteristics of children with developmental disabilities (DD) and typically developing (TD) children

The participants in this study included 131 children with developmental disabilities (DD) and 131 typically developing (TD) children. In the group of DD children, the proportion of males was more than twice that of females (Table 1).

TABLE I
Demographic characteristics of children with developmental disabilities (DD) and typically developing (TD) children

Demographic characteristics	DD Children (*n* = 131)	TD Children (*n* = 131)

Sex	Male	89	67.94	63	48.09
	Female	42	31.82	68	51.91

B. Comparison of developmental levels between children with DD and TD

Children with DD showed lower performance in all cognitive, language, social, and emotional tasks compared to TD children. These differences were all statistically significant (Figure 1). Especially in the cognitive and language domain, children with DD showed significant developmental delays compared to TD children in items such as asking questions like why and how (DD: 25.19%, TD: 94.66%, Diff: 69.47%), telling a story (DD: 13.74%, TD: 81.68%, Diff: 67.94%), asking who, what, when, and where (DD: 38.93%, TD: 96.95%, Diff: 58.02%), and explaining things (DD: 6.87%, TD: 57.25%, Diff: 50.38%). In the social and emotional domain, A developmental difference was observed in the item interest and curiosity, with children with DD showing delayed developmental levels compared to children with TD (DD: 31.31%, TD: 76.34%, Diff: 45.03%).

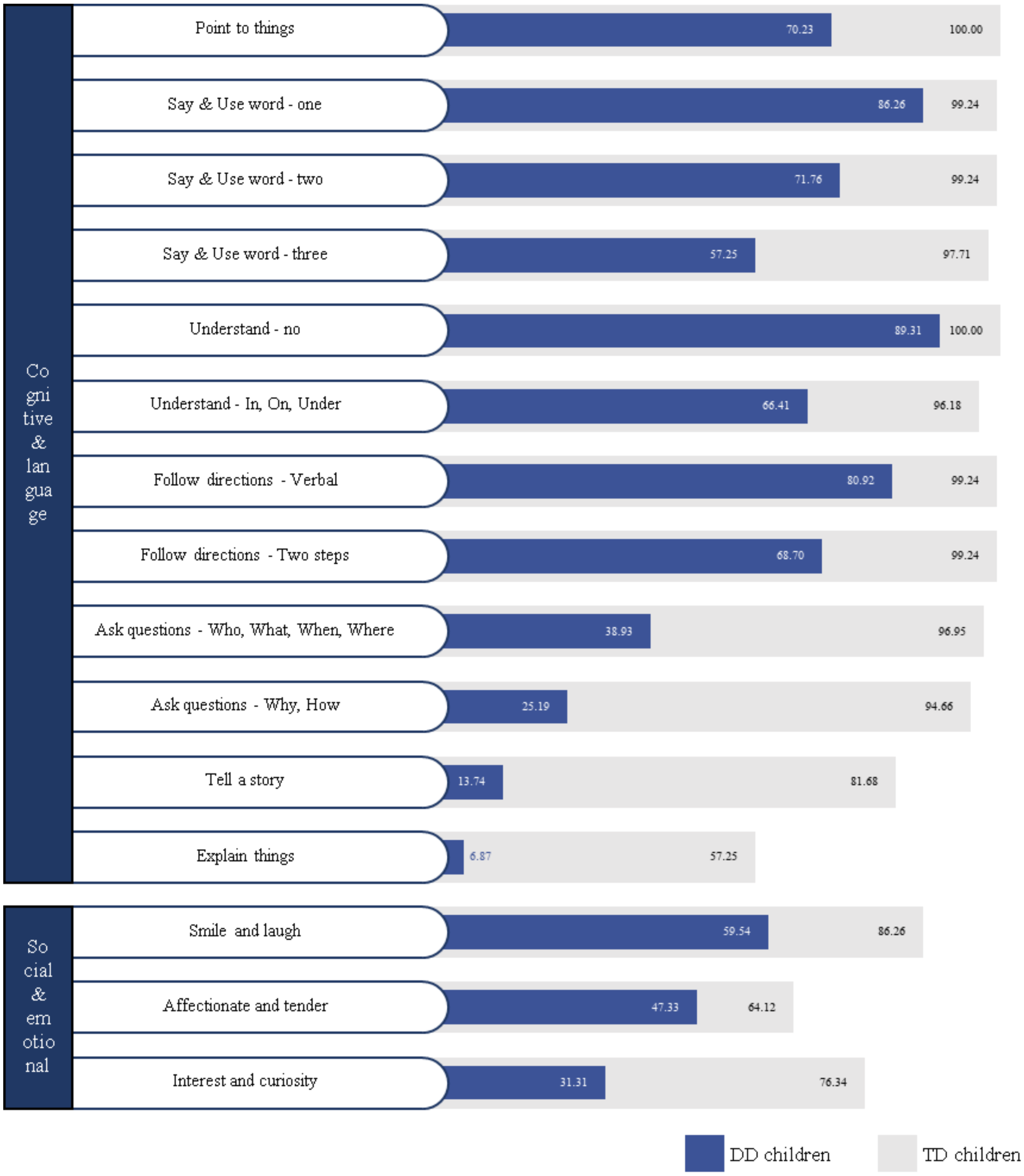


Figure 1. Comparison of developmental levels by domain between children with DD and TD

IV. Discussion

This study compared the developmental levels of cognitive & language, and social & emotional domains between three-year-old children with developmental disabilities (DD) and typically developing (TD) children to develop a virtual reality (VR) intervention aimed at enhancing the developmental levels of DD children. The study results showed that DD children exhibited significantly lower developmental levels in all domains and items compared to TD children. Based on these findings, it is recommended that the development of VR based interventions aimed at improving the developmental levels of children with DD include the following components.

According to the findings of this study, DD children in the cognitive and language domains compared to TD children, particularly in higher-level language tasks such as asking questions, telling stories, and explaining. This suggests difficulties not merely in vocabulary acquisition, but in the ability to express thoughts and understand situations through language [6], [7]. Therefore, VR intervention programs should provide children with repeated opportunities to express their thinking through language in virtual scenarios [3], [8], [9]. The content should include language-based cognitive training such as understanding story structures, cause-and-effect reasoning, and question-and-answer activities [10], [11]. Additionally, since DD children show less interest and curiosity for new things compared to TD children, incorporating familiar animals or characters they have previously experienced into program development can help them accept new concepts more comfortably [12], [13].

V. Conclusions

This study highlights the significant developmental differences between three-year-old children with developmental disabilities (DD) and typically developing (TD) children, particularly in cognitive & language and social & emotional domains. Based on these findings, VR intervention programs should target delayed developmental items in children with DD and provide tailored interventions that meet their specific needs.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176)

References

[1] X. Yang *et al.*, "Effectiveness of virtual reality technology interventions in improving the social skills of children and adolescents with autism: systematic review," *Journal of Medical Internet Research*, vol. 27, Art. no. e60845, 2025. [Online]. Available: https://doi.org/10.2196/60845

[2] J. C. O'Brien and H. M. Kuhaneck, Eds., *Case-Smith's Occupational Therapy for Children and Adolescents*, 8th ed. Elsevier, 2020.

[3] B. Karami, R. Koushki, F. Arabgol, M. Rahmani, and A. H. Vahabie, " Effectiveness of virtual/augmented reality–based therapeutic interventions on individuals with autism spectrum disorder: a comprehensive meta-analysis," *Frontiers in Psychiatry*, vol. 12, Art. no. 665326, 2021. [Online]. Available: https://doi.org/10.3389/fpsyt.2021.665326

[4] C. Li, M. Belter, J. Liu, and H. Lukosch, "Immersive virtual reality enabled interventions for autism spectrum disorder: A systematic review and meta-analysis," *Electronics*, vol. 12, no. 11, Art. no. 2497, 2023. [Online]. Available: https://doi.org/10.3390/electronics12112497

[5] S. C. Bodison, L. I. Stein Duker, B. Nakasuji, M. Gabriele, and E. I. Blanche, "Occupational therapy for children with autism spectrum disorder and intellectual and developmental disability," in *Handbook of Treatment Planning for Children with Autism and Other Neurodevelopmental Disorders*, Springer International Publishing, 2022, pp. 389–398.

[6] L. Shahmoradi and S. Rezayi, "Cognitive rehabilitation in people with autism spectrum disorder: a systematic review of emerging virtual reality-based approaches," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, no. 1, Art. no. 91, 2022. [Online]. Available: https://doi.org/10.1186/s12984-022-01069-5

[7] N. Marrus and L. Hall, "Intellectual disability and language disorder," *Child and Adolescent Psychiatric Clinics of North America*, vol. 26, no. 3 , pp. 539–552, 2017. [Online]. Available: https://doi.org/10.1016/j.chc.2017.03.001

[8] P. Mittal, M. Bhadania, N. Tondak, P. Ajmera, S. Yadav, A. Kukreti, *et al.*, "Effect of immersive virtual reality-based training on cognitive, social, and emotional skills in children and adolescents with autism spectrum disorder: A meta-analysis of randomized controlled trials," *Research in Developmental Disabilities*, vol. 151, Art. no. 104771, 2024. [Online]. Available: https://doi.org/10.1016/j.ridd.2024.104771

[9] F. Ke, J. Moon, and Z. Sokolikj, "Virtual reality–based social skills training for children with autism spectrum disorder," *Journal of Special Education Technology*, vol. 37, no. 1, pp. 49–62, 2020. [Online]. Available: https://doi.org/10.1177/0162643420945603

[10] L. J. Camilleri, K. Maras, and M. Brosnan, "Effective digital support for autism: Digital social stories," *Frontiers in Psychiatry*, vol. 14, Art. no . 1272157, 2024. [Online]. Available: https://doi.org/10.3389/fpsyt.2023.1272157

[11] S. L. Gillam, D. Hartzheim, B. Studenka, V. Simonsmeier, and R. Gillam, "Narrative intervention for children with autism spectrum disorder (ASD)," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 920–933, 2015. [Online]. Available: https://doi.org/10.1044/2015_JSLHR-L-14-0295

[12] N. Rakhymbayeva, A. Amirova, and A. Sandygulova, "A long-term engagement with a social robot for autism therapy," *Frontiers in Robotics and AI*, vol. 8, Art. no. 669972, 2021. [Online]. Available: https://doi.org/10.3389/frobt.2021.669972

[13] R. I. Gayle, A. L. Valentino, and A. M. Fuhrman, "Virtual reality training of safety and social communication skills in children with autism: an examination of acceptability, usability, and generalization," *Behavior Analysis in Practice*, vol. 17, pp. 1–17, 2024. [Online]. Available: https://doi.org/10.1007/s40617-024-00968-4

Design and Simulation of a 3D-Printed UHF Pyramidal Horn Antenna Using Dimensions Calculated by a Custom Tool

L. Saing[1], P. Kim[1, *], D. Bae[1], T. Thap[1]
Department of Telecommunication and Electronic Engineering, Royal University of Phnom Penh, Phnom Penh, Cambodia
*Contact: kim.phirun@rupp.edu.kh

Abstract— **In this paper, we introduce a calculation tool that is capable of calculating the dimension of the pyramidal horn antenna. This approach uses the Java programming language to process the design equation and software. The design of the antenna uses the dimension of the rectangular waveguide model WR510, and the calculation result is illustrated based on the simulated antenna in the HFSS simulation tool and the actual design using 3D-printed filament materialized by copper tape. The result from the calculation tool is used to align with the desired antenna that operates at the UHF band from 800 MHz to 2.5 GHz. From the simulation, the antenna was able to achieve a gain of 15 dB at a frequency of 1.75 GHz and a return loss value below -10 dB within a frequency range of 1.2 GHz to 2.3 GHz.**

I. Introduction

Nowadays, the design of various types of antennas is very crucial, as they are intended to be more effective in both cost and performance. Those antennas need to be flexible in order to be integrated into many devices, systems, and applications. Among them, horn antennas are widely used in many applications, such as microwave communication, radar systems, radio astronomy, wireless communication, and more, as they have positive impacts with high gain, high directivity, wide bandwidth, and a low SWR (standby wave ratio).

We derived the fundamental equation for antenna design from Balanis (2016) [1], which calculates the aperture dimension. Y. He et al. [2] presented a broadband double-ridged horn antenna (DRHA) for the 5G millimeter-wave band. Yao et al. [3] proposed the design and fabrication of a Ka-band antenna using 3D printing technology. Lomakin et al. [4] investigate the design optimization of a pyramidal horn antenna for 3D printing in the millimeter-wave range. Helena et al. [5] showed two 3D-printed antennas for satellite services in which the first one used copper tape on plastic while the other used conductive plastic. It showed that the copper tape antenna was more efficient but didn’t have better bandwidth. X. Qing, T. S. Tan, and Z. N. Chen (2009) [6] created a calculation tool based on the transmission line model for the H-shaped patch antenna. M. C. de Melo et al. (2021) [7] proposed a computational intelligence-based methodology for antenna design using a surrogate model and multi-objective optimization algorithms. L. Linkous, J. Lundquist, and E. Topsakal (2023) [8] introduced an open-source graphical tool, AntennaCAT, for automating antenna design, calculation, and CAD creation. The tool is also capable of supporting a number of commercial simulation software.

To design a pyramidal horn antenna, various detailed calculation processes need to be used to obtain the essential parameters that can provide optimum results. Initially, the use of a normal calculator with a long equation and many parameters was time-consuming and prone to errors. It can be resolved with the assistance of a software tool that can quickly and accurately calculate the dimension of the pyramidal horn antenna. It minimizes numerical errors while requiring only a few input parameters. Additionally, the dimension of the antenna, especially the flare angles and the aperture dimension, is very crucial in determining the performance. To ease this issue, the antenna was designed and fabricated using 3D-printed material. The design uses Fusion 360, a software tool that creates 3D objects and provides detailed dimensions to construct the structure. Moreover, achieving optimal antenna performance may require expensive materials, especially with thicker sheets like brass, copper, or aluminum, to minimize loss. The fabrication of the antenna is also a complicated process. The frequency operation also influences the antenna's cost. This is because there are some antennas that can operate at lower frequencies; thus, the antenna becomes bigger and requires a large amount of material. With the use of 3D-printed material and copper tape, it ensures both cost-effective and reliable performance. This structure allowed for quick and cheap prototype antenna development.

The objective of this paper is the introduction of a software tool for calculating the dimensions of the pyramidal horn antenna. The calculated dimension is simulated in HFSS while the structure of the antenna is constructed using 3D-printed filament. It was later materialized using copper tape and testing in the experiment.

II. METHOD

A. Rectangular Waveguide

To efficiently transmit electromagnetic energy or power from one point in space to another, rectangular waveguides are utilized [9]. The dimension of the rectangular waveguide is crucial for the performance of the whole antenna structure. In Fig. 2, various parameters such as *a*, b, L_1, *h*, L_2 , and *a/2* are important in the design of the rectangular waveguide and its coaxial probe feed. In this design, the proposed antenna operates in the desired frequency range from 800 MHz to 2.5 GHz. The rectangular waveguide has the center frequency set at *f*=1.75 GHz, and the standard model for this frequency is WR510.

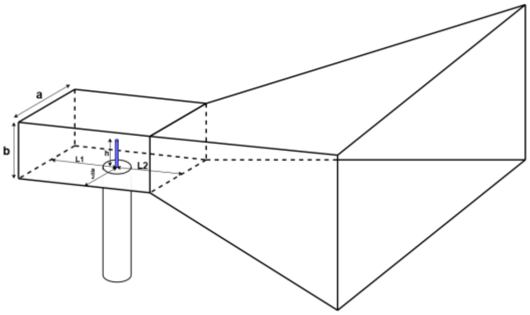


Fig. 1. Structure of the rectangular waveguide

From [10], it has a dimension of 129.54 × 64.77 mm, and the larger and smaller dimensions are presented by a and *b*, respectively. This model operates in a frequency range from 1.45 GHz to 2.2 GHz, which is suitable for the desired bandwidth. The other parameters, such as cutoff frequency for both the lowest order and the next mode, are shown in Table I.

TABLE I
Dimension of the Rectangular Waveguide Model WR510

a, Width of the rectangular waveguide	129.54mm
b, Height of the rectangular waveguide	64.77mm
Bandwidth	1.45GHz-2.2GHz
Cutoff frequency lowest order mode	1.157GHz
Cutoff frequency next mode	2.314GHz

$$\lambda_g = \frac{\lambda_0}{\sqrt{1-\left(\frac{\lambda_0}{\lambda_c}\right)^2}} \tag{1}$$

$$\lambda_0 = \frac{C}{f} \tag{2}$$

$$\lambda_c = \frac{c}{f_c} \tag{3}$$

$$f_c = \frac{C}{2a} \tag{4}$$

$$L_1 = \frac{3\lambda_g}{4} \tag{5}$$

$$h = \frac{\lambda_0}{4} \tag{6}$$

$$L2 = \frac{\lambda_g}{4} \tag{7}$$

To calculate the position and dimension of the feed of the rectangular waveguide, dimensions a and b of the selected model are used. From [11], λ_g , the waveguided wavelength is calculated using (1), λ_0 where is the free-space wavelength and λ_c is the cutoff wavelength. To calculate free-space wavelength and cutoff wavelength, we use (2) and (3), respectively. L_1 is the distance from the waveguide opening to the backshort in a rectangular waveguide, which is calculated using (5). The height of the pin feed, *h*, and the probe-to-aperture distance, L_2, are calculated using (6) and (7), respectively.

B. Aperture of the Antenna

Apart from the rectangular waveguide, the aperture can be explained as the open end of the horn through which electromagnetic waves are radiated or received. Connected to the rectangular waveguide, it determines the gain, radiation pattern, and efficiency of the antenna, making it a crucial part of the antenna.

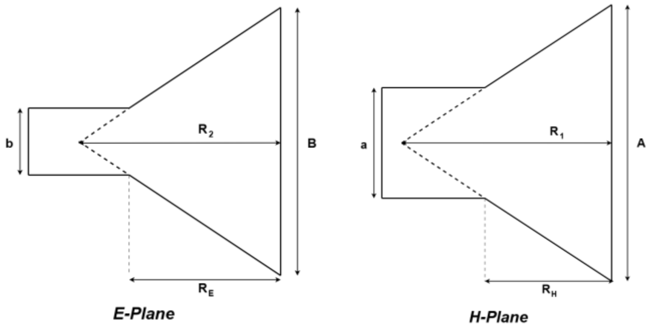


Fig. 2. E-Plane and H-Plane of the antenna

In this design, the aperture dimension of the pyramidal horn antenna is calculated using (8). The desired gain is set at 15 dB at *f* = 1.75 GHz, and later the value for dimensions *A* and *B* of the aperture is determined from (9) and (10). From [12], parameter *K* is the coefficient that determines the approximated value of the aperture's width. This parameter can be determined when the condition that $R_E = R_H$ from (11) and (12) below is satisfied.

$$A = k\lambda\sqrt{G} \tag{9}$$

$$B = \frac{1}{4\pi} \times \frac{G\lambda_0{}^2}{0.51 \times a_1} \tag{10}$$

$$R_E = R_2\left(1-\frac{a}{A}\right) \tag{11}$$

$$R_H = R_1\left(1-\frac{b}{B}\right) \tag{12}$$

C. Software Design

In the proposed system, the software is developed using the Java programming language combined with the described parameters and equation. This software tool operates as a calculator, accepting inputs such as *a*, *b*, desired gain, and center frequency. The coefficient *K* determines the satisfaction level between R_E and R_H in the equations at (11) and (12). As shown in fig. 4, this software generates the coefficient, which is later applied to the equation. It adjusted to minimize the difference between the target and computed results. To ensure numerical precision, the value of *K* is represented using the double data type, which provides higher accuracy compared to float. All parameter values, which are the output from the process—specifically (1), (2), (3), (4), (5), (6), (7), (9), (10), (11), and (12)—are expressed in millimeters. These values are used in the simulation, design, and fabrication of the antenna, which served as an illustration of the proposed software, as shown in Fig. 4.

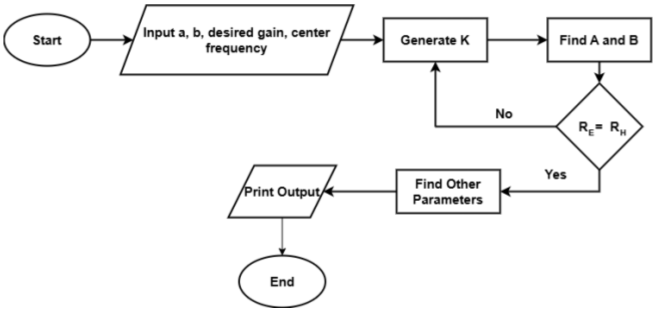
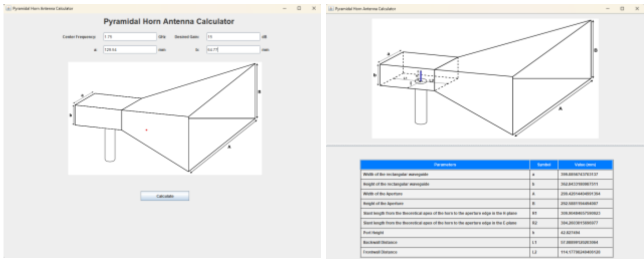


Fig. 3. Overview of software computation process


Pyramidal Horn Antenna Calculator
Pyramidal Horn Antenna Calculator
Center Frequency:
GHz
Desired Gain:
dB
a:
mm
b:
mm
a
b
A
B
Calculate

Center Frequency: 1.75 GHz Desired Gain: 15 dB
a: 129.54 mm b: 64.77 mm

Parameters	Symbol	Value (mm)
Width of the rectangular waveguide	a	[illegible]
Height of the rectangular waveguide	b	[illegible]
Width of the Aperture	A	[illegible]
Height of the Aperture	B	[illegible]
Slant length from the theoretical apex of the horn to the aperture edge in the H-plane	R1	[illegible]
Slant length from the theoretical apex of the horn to the aperture edge in the E-plane	R2	[illegible]
Post Height	h	[illegible]
Backwall Distance	L1	[illegible]
Frontwall Distance	L2	[illegible]

Fig. 4. Proposed software functions as a calculator

III. Simulation

A. Design Setup

In this simulation, HFSS, a powerful simulation software, is used to design, simulate, and analyze the performance of the pyramidal horn antenna. The antenna is 10 mm in thickness and uses copper as its material while being put in the air as its environment. It allows antennas to accurately model how electromagnetic waves radiate and interact with the surrounding environment. It is designed with an impedance of 50 Ω to ensure optimal matching with standard RF systems.

B. Simulation Results

In this simulation, return loss, VSWR, gain, and radiation patterns are presented. The results from this simulation assist in evaluating the effectiveness of the design for the intended frequency band. Moreover, they serve as illustrations of the proposed method described above with the parameter calculated by the software tool.

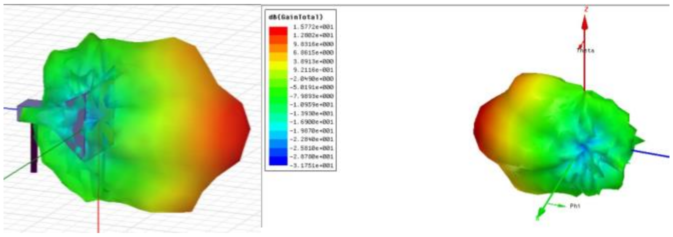


Fig. 5. 3D radiation pattern of the pyramidal horn antenna

In Fig. 5, the gain of the simulated antenna showed that the antenna is highly directional, with the strongest signal around 15.72 dB gain. Besides, it significantly reduces its signal strength in all other directions to -31.24 dB. The 2D plot of radiation from Fig. 6 ensured that the antenna radiates predominantly in the direction of 180 degrees. Nonetheless, it has much weaker radiation in other directions, such as side lobes and back lobes. Initially, the antenna showed almost no radiation on either side of the main lobe.

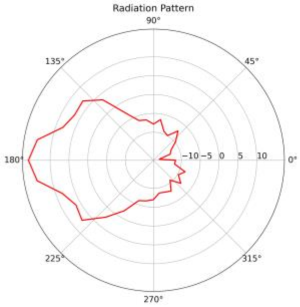


Fig.6. Radiation pattern of the designed antenna

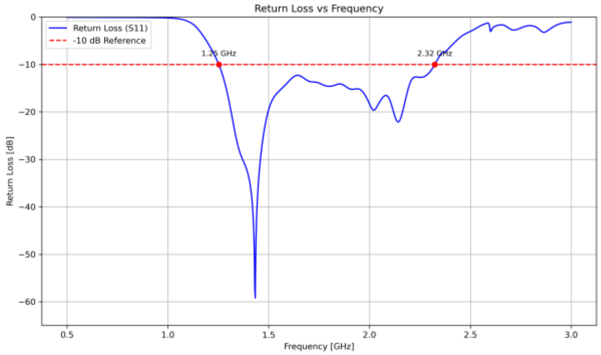


Fig. 7. Return loss of the pyramidal horn antenna

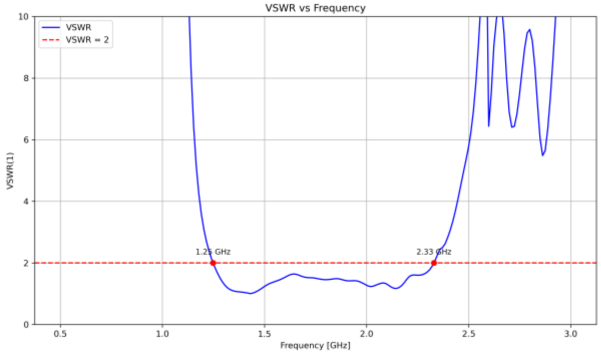


Fig. 8. VSWR of the designed antenna

Fig. 7 shows that the return loss is lower than -10 dB within a frequency range of 1.2 GHz to 2.3 GHz, indicating that the antenna is effectively matched to its transmission line over that range. This ensures efficient power transfer and minimal reflections. Fig. 5 shows the resonant frequency at 1.4GHz where the return loss of the antenna, $|S_{11}| = -51.57dB$, proves that it is near-perfect impedance matching. The return loss value at center frequency *f*=1.75GHz. From Fig. 6, the result of the simulation showed that the designed antenna maintains a VSWR below 2 between 1.25 GHz and 2.329 GHz. This implies that it has a good impedance matching and efficient transmission of power over the above-mentioned band. The antenna gives a virtually 1.5036 VSWR at the frequency at the center. The 2D plot of radiation from Fig. 7 ensured that the antenna radiates predominantly in the direction of 180 degrees. Nonetheless, it has much weaker radiation in other directions, such as side lobes and back lobes. Initially, the antenna showed almost no radiation on either side of the main lobe.

IV. Fabrication and MEASUREMENT

A. Fabrication Process

The antenna was designed in Fusion 360 using the same calculated dimension as the dimension of the simulated antenna. The 3D-designed antenna was printed and constructed into two parts using 3D-printed filament, ensuring precise dimensions as shown in Fig. 9. Later, the 3D-printed antenna was materialized and connected using copper tape.

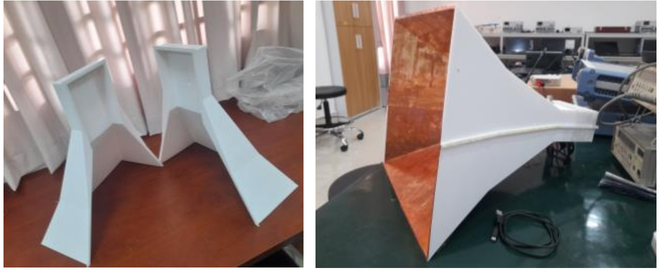


Fig. 9 Pyramidal horn antenna before (left) and after (right) materialization

B. Experiment Setup and Result

The return loss experiment was conducted by connecting the fabricated antenna to a calibrated VNA. The setup measures the S11 parameter with a coaxial cable from the LibreVNA kit. Calibration was done using an SOL (short, open, load) kit. The return loss of the antenna was recorded across a defined frequency range.

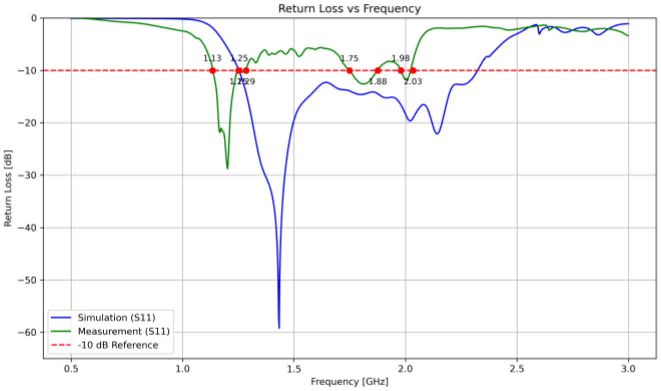


Fig. 10. Comparison of simulated and measured return loss for the designed antenna

From Fig. 10, the return loss from the experiment was used to compare with the return loss from the simulation of the designed antennas. The graph indicates resonant frequencies around 1.4 GHz and 2.1 GHz, where impedance matching is optimized. Besides this, the simulation accurately predicts these resonant frequencies. However, the measured return loss is higher than simulated, which indicates greater reflected power. This may happen due to impedance mismatches in the fabricated pyramidal horn antenna.

V. Conclusions

In this paper, a 3D-printed UHF pyramidal horn antenna operating from 1.2 GHz to 2.3 GHz is designed, simulated, fabricated, and measured. The design of the antenna uses the dimension calculated by a custom calculation tool developed using the Java programming language. In addition, the dimension was used to design a 3D structure of the antenna, printed using 3D-printed filament, and fabricated with copper tape. The performance of the antenna was able to achieve a great gain with acceptable return loss and almost 1 GHz bandwidth, according to the simulation. Despite this, the experiment produced higher return loss and low gain due to impedance mismatches in the fabricated pyramidal horn antenna and the testing environment.

References

[1] C. A. Balanis, Antenna theory: *Analysis and design*, 4 th ed. Hoboken, NJ, USA: Wiley, 2016.

[2] Y. He et al., "Design of Broadband Double-Ridge Horn Antenna for Millimeter-Wave Applications," *IEEE Access*, vol. 9, pp. 118919-118926, 2021.

[3] H. Yao, S. Sharma, R. Henderson, S. Ashrafi and D. MacFarlane, "Ka band 3D printed horn antennas," *Texas Symposium on Wireless and Microwave Circuits and Systems (WMCS)*, Waco, TX, USA, 2017. doi: 10.1109/WMCaS.2017.8070701.

[4] K. Lomakin, J. Schür and G. Gold, "Design Optimization of Pyramidal Horn Antennas for 3D Printing in the mm-Wave Range," *16th European Conference on Antennas and Propagation (EuCAP)*, Madrid, Spain, 2022.

[5] D. Helena, A. Ramos, T. Varum and J. N. Matos, "Evaluation of Different Materials to Design 3D Printed Horn Antennas for Ku-Band," *2019 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)*, Aveiro, Portugal, 2019, pp. 1-3, doi: 10.1109/IMOC43827.2019.9317424.

[6] X. Qing, T. S. Tan, and Z. N. Chen, "Development of efficient calculation tool for miniaturized H-shaped antenna for UHF RFID applications," *2009 Asia Pacific Microwave Conference*, Singapore, 2009, pp. 2236-2239, doi: 10.1109/APMC.2009.5385426.

[7] M. C. de Melo, P. B. Santos, E. Faustino Jr., C. J. A. Bastos-Filho, and A. C. Sodré Jr., "Computational intelligence-based methodology for antenna development," *IEEE Access*, vol. 9, pp. 194276-194287, Nov. 2021, doi: 10.1109/ACCESS.2021.3137198.

[8] L. Linkous, J. Lundquist, and E. Topsakal, "AntennaCAT: Automated Antenna Design and Tuning Tool," *2023 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium)*, Portland, OR, USA, 2023, pp. 89-90, doi: 10.23919/USNC-URSI54200.2023.10289238.

[9] N. S. S. Saindla, A. K. Yellola, S. Sabavath, N. K. Uppari and M. Basha, "Design and study of waveguide using HFSS-High Frequency Structural Simulator," *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, Tirunelveli, India, 2017, pp. 284-286, doi: 10.1109/ICOEI.2017.8300933.

[10] everything RF, “Waveguide sizes | Dimensions & cutoff frequency.” Accessed: Mar. 18, 2025. [Online]. Available: https://www.everythingrf.com/tech-resources/waveguides-sizes.

[11] Y. Jang, "Design of a broadband double-ridged horn antenna," M.S. thesis, California State University, Los Angeles, CA, USA, 2024. [Online]. Available: https://scholarworks.calstate.edu/concern/theses/x059cf806

[12] M. B. R. Murthy, M. Sudhakar, and L. Bhogadi, "Design and testing of pyramidal horn," IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), vol. 10, pp. 79–85, 2015, doi: 10.9790/2834-10327985.

Development of Standard Data Management Platform for Sarcopenia Data Collection

SiHyeong Noh[1], Jin-Gyeong Lee[1], DongWook Lim[1], Go-eun Lee[1], Hee-Kyung Moon[2], Ji Hee Kim[3],

Seong-Kyu Choe[1], Chang-Won Jeong [1,4,*]

[1] STSC Center, Wonkwang University, Iksan 54538, Republic of Korea

[2] Institute for Educational Innovation, Wonkwang University, Iksan 54538, Republic of Korea

[3]Department of Rehabilitation Medicine, Wonkwang University School of Medicine and Hospital

[4] Smart Team, Wonkwang University Hospital, 54538, Republic of Korea

*Contact: mediblue@wku.ac.kr, phone +82-10 3674 2977

Abstract— **This study is a data management platform for the diagnosis of sarcopenia, and a web-based platform was developed to diagnose sarcopenia through physical function evaluation and collect clinical information related to it along with the establishment of a sarcopenia diagnosis process in the clinical field. Sarcopenia is characterized by a decrease in muscle mass and physical function, and early detection is important. Therefore, the EWGSOP and AWGS have provided guidelines to diagnose sarcopenia through various methods in clinical settings. Although efforts are being made to apply these guidelines in clinical practice, there is still a lack of integrated management. To this end, data can be systematically collected and managed through the platform and provided as utilization data for various studies on classified sarcopenia patients.**

I. Introduction

Sarcopenia is a disease related to muscle mass loss caused by various factors such as aging, chronic diseases, and decreased physical activity, and is emerging as a serious medical problem due to poor physical function and quality of life. In the management of sarcopenia, a standardized data management and an integrated platform are essential to maintain consistency between diagnosis and treatment. However, currently, there is a problem that it is difficult to integrate and compare data in clinical research and practice because the criteria and data management methods related to the diagnosis of sarcopenia differ from study to study. In this paper, based on the diagnostic criteria for sarcopenia and the definition of data obtained from patients in the clinical field, sarcopenia data were standardized and developed as a web-based platform. It was created for medical staff to use inside the hospital support for diagnosis sarcopenia, and access from outside is blocked. After that, through the proposed platform design and implementation plan, a standardized data management system is established, and the possibility of application to sarcopenia management and clinical sites in the future is suggested[1,2].

.

II. METHOD

Figure 1 shows the software structure of the web-based data management platform that we developed. Images and data are stored and managed as files on node servers that allow JavaScript code to be executed outside of the browser through Nginx, a web application server (WAS), and mongoDB, which is strong in handling large amounts of data, was used as the database.

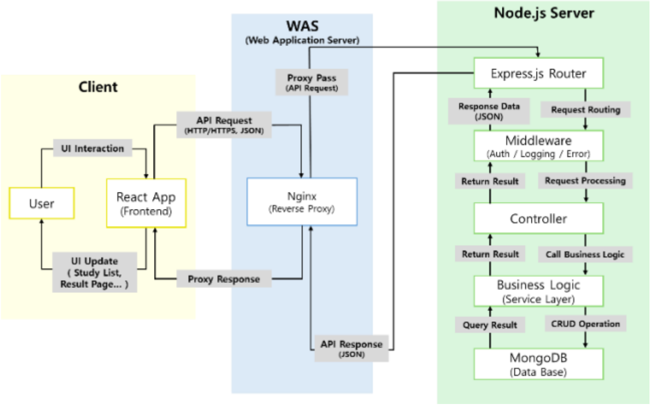
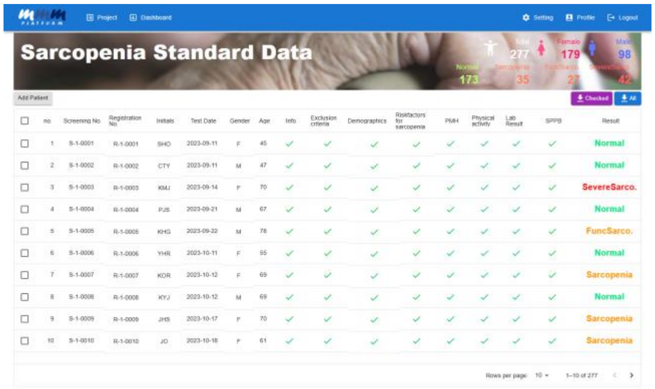


Fig. 1 Software structure for web-based data collection management



☐	no.	Screening No.	Registration No.	Initials	Test Date	Gender	Age	Info	Exclusion criteria	Demographics	Riskfactors for sarcopenia	PMH	Physical activity	Lab Result	SPPB	Result
☐	1	S-1-0001	R-1-0001	SHO	2023-09-11	F	45	✓	✓	✓	✓	✓	✓	✓	✓	Normal
☐	2	S-1-0002	R-1-0002	CTY	2023-09-11	M	47	✓	✓	✓	✓	✓	✓	✓	✓	Normal
☐	3	S-1-0003	R-1-0003	KMJ	2023-09-14	F	70	✓	✓	✓	✓	✓	✓	✓	✓	SevereSarco.
☐	4	S-1-0004	R-1-0004	PJS	2023-09-21	M	67	✓	✓	✓	✓	✓	✓	✓	✓	Normal
☐	5	S-1-0005	R-1-0005	KHG	2023-09-22	M	78	✓	✓	✓	✓	✓	✓	✓	✓	FuncSarco.
☐	6	S-1-0006	R-1-0006	YHR	2023-10-11	F	55	✓	✓	✓	✓	✓	✓	✓	✓	Normal
☐	7	S-1-0007	R-1-0007	KOR	2023-10-12	F	69	✓	✓	✓	✓	✓	✓	✓	✓	Sarcopenia
☐	8	S-1-0008	R-1-0008	KYJ	2023-10-12	M	69	✓	✓	✓	✓	✓	✓	✓	✓	Normal
☐	9	S-1-0009	R-1-0009	JHS	2023-10-17	F	70	✓	✓	✓	✓	✓	✓	✓	✓	Sarcopenia
☐	10	S-1-0010	R-1-0010	JO	2023-10-18	F	61	✓	✓	✓	✓	✓	✓	✓	✓	Sarcopenia

Rows per page: 10 ▾ 1–10 of 277 ‹ ›

Fig. 2 web-based data collection management platform

Figure 2 is the result screen of the collection of standard data for web-based sarcopenia, of which the total number of people collected is 277, of which 179 are female and 98 are male. And the normal number was 173 and there were 35 patients with sarcopenia, 27 patients with functional sarcopenia, and 42 patients with severe sarcopenia, and 104 patients were identified as sarcopenia. The collected data were collected in accordance with the IRB's approval(IRB No. WKUH 2023-08-031).

The data we want to collect is demographic characteristics, risk factors for sarcopenia, drug history, physical activity, Lab results and physical function evaluation. Sarcopenia patients were collected by classifying the results of sarcopenia diagnosis into normal, sarcopenia, severe sarcopenia using the Asian Classification Scale (AWGS) through the evaluation of muscle mass, grip strength, and physical function. And having low muscle strength with low physical performance is considered clinically relevant and newly defined as “functional sarcopenia”[3]. Related data were collected by the hospital's Department of Rehabilitation Medicine.

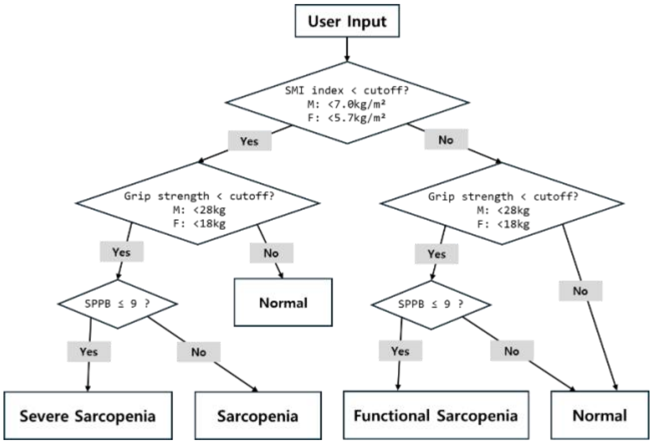


Fig. 3 Process of diagnosing sarcopenia

As a method of diagnosing sarcopenia in the clinical field, the process of evaluating muscle mass, grip strength, and physical function is shown in Figure 3. In this regard, if the grip strength, muscle mass index, and SPPB score all fail to meet specific criteria, it is diagnosed as "severe sarcopenia," and among them, if the grip strength or SPPB score is insufficient, it is diagnosed as pre- sarcopenia, and if the grip strength and muscle mass index or SPPB score are insufficient, it is diagnosed as sarcopenia. Detailed criteria for diagnosis are as follows[4].

1. SPPB (≤9 points)

‣Static balance (4 points)

- Whether to maintain the general position for 10 seconds (1 point)
- Hold the anti-aligned position for 10 seconds (1 point)
- Whether to maintain the line position for 10 seconds (2 points)

‣Walking speed (4 points)

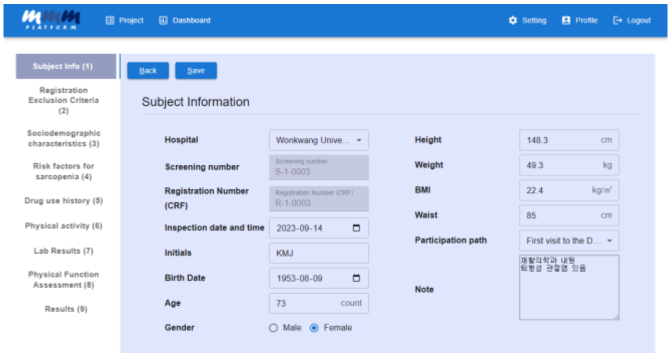
- Time and speed required to walk 4 m (m/s)

‣Standing up (4 points)

- Time required to get up 5 times from the chair

III. Result

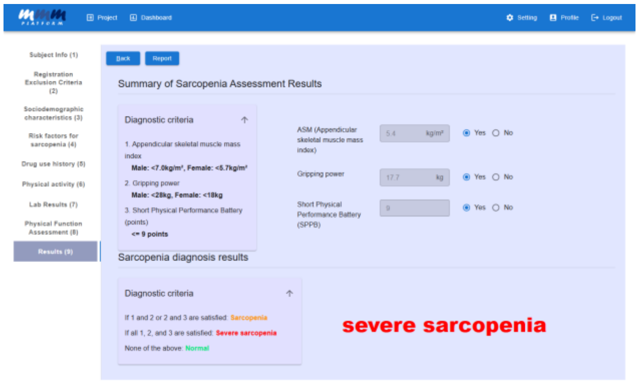
To collect the standard data of the study subjects as shown in Figure 4, data on demographic characteristics, risk factors for sarcopenia, drug history, physical activity, and laboratory outcomes were collected in addition to the diagnostic measurements mentioned above.



Subject Information

Hospital	Wonkwang Unive...	Height	148.3 cm
Screening number	S-1-0003	Weight	49.3 kg
Registration Number (CRF)	R-1-0003	BMI	22.4 kg/m²
Inspection date and time	2023-09-14	Waist	85 cm
Initials	KMJ	Participation path	First visit to the D...
Birth Date	1953-08-09	Note	재활의학과 내원 퇴행성 관절염 있음
Age	73 count		
Gender	○ Male ● Female		

Fig. 4 Standard sarcopenia data Study subject information



Summary of Sarcopenia Assessment Results

Diagnostic criteria

1. Appendicular skeletal muscle mass index
   Male: <7.0kg/m², Female: <5.7kg/m²
2. Gripping power
   Male: <28kg, Female: <18kg
3. Short Physical Performance Battery (points)
   <= 9 points

ASM (Appendicular skeletal muscle mass index)	5.4 kg/m²	● Yes ○ No
Gripping power	17.7 kg	● Yes ○ No
Short Physical Performance Battery (SPPB)	9	● Yes ○ No

Sarcopenia diagnosis results

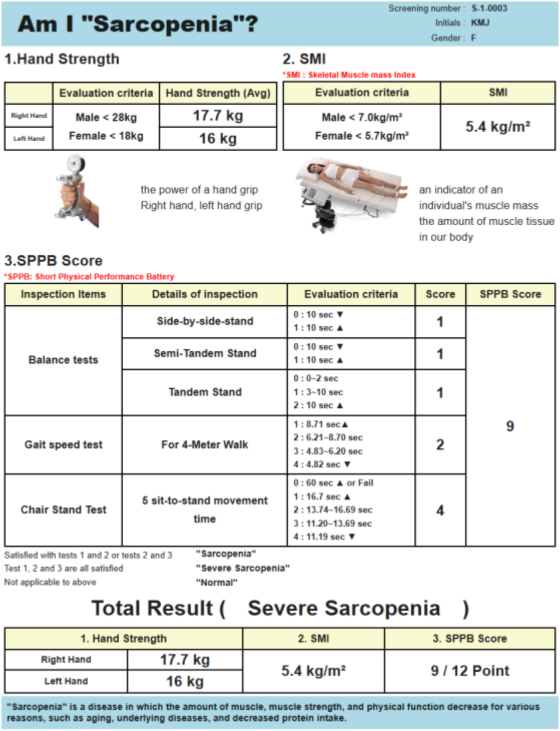
Diagnostic criteria

If 1 and 2 or 2 and 3 are satisfied: Sarcopenia
If all 1, 2, and 3 are satisfied: Severe sarcopenia
None of the above: Normal

severe sarcopenia

Fig. 5 sarcopenia diagnosis result screen

The results of the study's diagnosis of sarcopenia were confirmed through the results measured in the clinical field as shown in Figure 5. It can be confirmed that the study subjects displayed on the result screen were judged as 'severe sarcopenia' because they did not meet the evaluation criteria as a result of evaluating the grip strength, muscle mass index, and physical function.



Am I "Sarcopenia"?

Screening number : S-1-0003
Initials : KMJ
Gender : F

1.Hand Strength

	Evaluation criteria	Hand Strength (Avg)
Right Hand	Male < 28kg Female < 18kg	17.7 kg
Left Hand		16 kg

the power of a hand grip
Right hand, left hand grip

2. SMI

*SMI : Skeletal Muscle mass Index

Evaluation criteria	SMI
Male < 7.0kg/m² Female < 5.7kg/m²	5.4 kg/m²

an indicator of an individual's muscle mass the amount of muscle tissue in our body

3.SPPB Score

*SPPB: Short Physical Performance Battery

Inspection Items	Details of inspection	Evaluation criteria	Score	SPPB Score
Balance tests	Side-by-side-stand	0 : 10 sec ▼ 1 : 10 sec ▲	1	9
	Semi-Tandem Stand	0 : 10 sec ▼ 1 : 10 sec ▲	1	
	Tandem Stand	0 : 0~2 sec 1 : 3~10 sec 2 : 10 sec ▲	1	
Gait speed test	For 4-Meter Walk	1 : 8.71 sec ▲ 2 : 6.21~8.70 sec 3 : 4.83~6.20 sec 4 : 4.82 sec ▼	2	
Chair Stand Test	5 sit-to-stand movement time	0 : 60 sec ▲ or Fail 1 : 16.7 sec ▲ 2 : 13.74~16.69 sec 3 : 11.20~13.69 sec 4 : 11.19 sec ▼	4	

Satisfied with tests 1 and 2 or tests 2 and 3	"Sarcopenia"
Test 1, 2 and 3 are all satisfied	"Severe Sarcopenia"
Not applicable to above	"Normal"

Total Result (Severe Sarcopenia)

1. Hand Strength		2. SMI	3. SPPB Score
Right Hand	17.7 kg	5.4 kg/m²	9 / 12 Point
Left Hand	16 kg		

"Sarcopenia" is a disease in which the amount of muscle, muscle strength, and physical function decrease for various reasons, such as aging, underlying diseases, and decreased protein intake.

Fig. 5 Sarcopenia diagnostic results report

The result report in Figure 5 makes the diagnosis result of sarcopenia easier to understand. It is organized so that the patient can easily recognize the diagnosis result through figures and tables. In clinical settings, this report is provided to patients.

IV. Conclusion

We proposed a standard data management platform for sarcopenia to collect various data on subjects of sarcopenia disease. In particular, it is measured in the clinical field to confirm the results, and based on this, it is used to present guidelines for exercise prescriptions and diet as well as awareness of sarcopenia to patients. As a future study, we plan to conduct research to develop information on exercise prescriptions and diets and services linked to the local food industry. And we will also increase the number of samples by collecting more patient data.

Acknowledgment

This study was supported by the grant of the National Research Foundation of Korea funded by the Korean government (MSIP) (No. RS2021NR060106) and supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare (RS-2024-00439936).

References

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] J.Y Baek, H.W Jung, K.M Kim, M.J Kim, C.Y.J Park, K.P Lee, S.Y Lee, I.Y Jang, O.H Jeon, J.Y Lim(2023), Korean Working Group on Sarcopenia Guideline: Expert Consensus on Sarcopenia Screening and Diagnosis by the Korean Society of Sarcopenia, the Korean Society for Bone and Mineral Research, and the Korean Geriatrics Society, Ann Geriatr Med Res, 27(1), 9-21. https://doi.org/10.4235/agmr.23.0009

[4] Won Chang Won. (2020). Diagnosis of sarcopenia in primary health care. J Korean Med Assoc, 63(10), 633–641. https://doi.org/10.5124/jkma.2020.63.10.633

Blendshape-Guided 3D Facial Image and Emotion Recognition Framework Using Meta Quest Pro Data

Hyeonwoo Kim
Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea
*Contact: hwkim24@sch.ac.kr

Abstract— **We propose a novel framework that leverages blendshape coefficients obtained from Meta Quest Pro to generate expressive 3D facial images and utilize them for robust emotion recognition. Our approach integrates real-time facial tracking data, 3D avatar rendering through Blender, and multimodal deep learning techniques combining both rendered facial images and blendshape vectors. By introducing a synthetic data generation mechanism that mirrors actual user expressions in VR, this framework paves the way for scalable and personalized emotion recognition systems. While this study focuses on the conceptual design of the system, we outline the full technical pipeline and discuss the potential applications and challenges of this approach, laying a solid foundation for future empirical studies.**

I. Introduction

Emotion recognition is an essential component of affective computing, with applications spanning virtual reality (VR), education, healthcare, gaming, and human-computer interaction (HCI) [1]. In particular, the ability to detect user emotions in immersive environments enhances user experience and enables intelligent adaptation of content. Recent advances in facial tracking devices, such as Meta Quest Pro, provide new opportunities to capture high-fidelity facial expression data in the form of blendshape coefficients [2].

Blendshape coefficients represent the activation levels of predefined facial muscle movements and are widely used in facial animation and avatar-driven systems [3]. However, the use of this data for emotion recognition has remained underexplored, especially in contexts where privacy, realism, and multimodal learning are paramount. Traditional facial expression datasets often rely on 2D images with limited variation and demographic representation, whereas blendshape-based representations offer a structured, low-dimensional, and privacy-preserving alternative [4].

Facial expression recognition (FER) has traditionally been approached through the analysis of static 2D images using convolutional neural networks (CNNs) or recurrent models for video-based inputs. Large-scale datasets such as FER2013, AffectNet, and CK+ have fueled the development of deep FER models [5]. However, these datasets often suffer from bias, lack of 3D information, and limited context regarding the user's environment, particularly in VR settings. Meanwhile, blendshapes have been a standard tool in 3D animation pipelines, enabling realistic facial deformation through linear combinations of predefined facial poses. In VR and AR, devices like Meta Quest Pro expose these coefficients for real-time tracking, enabling avatar-driven interaction. Prior research has utilized blendshapes for avatar control and speech animation [6], but few studies have attempted to use them for emotion recognition or analysis. Moreover, multimodal emotion recognition research has primarily focused on fusing audio, text, and visual data, while fusion involving 3D facial geometry and rendered appearance features is relatively underexplored [7].

In this paper, we propose a novel framework that bridges real-time blendshape tracking and 3D facial rendering to facilitate emotion recognition. Our method not only uses the blendshape coefficients themselves but also renders corresponding 3D facial images through Blender, enabling multimodal training with both geometric and appearance features. The rendered facial expressions reflect user-specific nuances and are fully controllable, allowing for fine-grained emotion analysis.

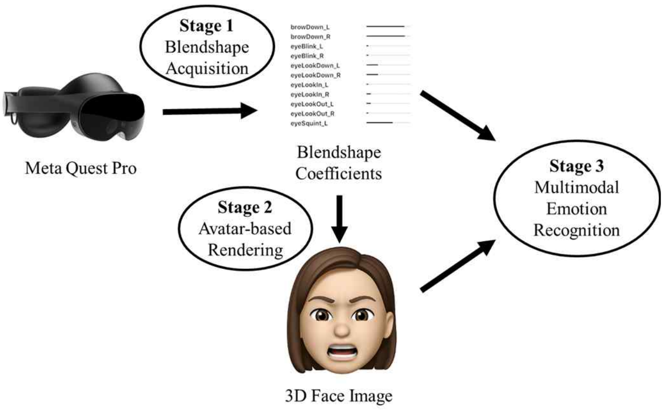


Fig. 1 Overview of the proposed framework

II. Proposed Framework

Fig. 1 shows overview of our proposed framework. Our framework comprises three main stages: (1) blendshape acquisition, (2) avatar-based 3D facial image rendering, and (3) multimodal emotion recognition. The architecture is modular and designed for scalability and real-time application.

A. Stage 1: Blendshape Acquisition

Using Meta Quest Pro, we collect blendshape coefficient vectors that describe facial muscle movements. These vectors are timestamped and recorded at a high frame rate, offering dynamic tracking capabilities. Each vector includes values corresponding to expressions such as smiling, frowning, eyebrow raises, and mouth movements. The data is stored as CSV or JSON files and used as ground truth for rendering and modeling.

B. Stage 2: Avatar-Based Rendering via Blender

We utilize a 3D avatar (e.g., Ready Player Me models) embedded with shape keys that correspond to blendshape parameters. A Python script within Blender applies blendshape values to these shape keys to simulate expressions. The scene is automatically configured with lighting, camera, and rendering settings to output realistic 3D facial images. These images mirror the expression encoded by each blendshape vector, producing a consistent and customizable visual dataset.

C. Stage 3: Multimodal Emotion Recognition

The core model ingests both the rendered 3D facial image and its associated blendshape vector. The image is processed through a CNN backbone (e.g., ResNet or EfficientNet), while the blendshape vector is passed through a multi-layer perceptron (MLP). The extracted features are concatenated and fed into a classification head trained to predict one of several predefined emotional states (e.g., happy, sad, surprised, angry, neutral). This fusion enables the model to learn complementary visual and geometric features, improving generalizability and robustness.

III. Potential Applications and Advantages

The proposed framework offers several key advantages. First, the use of blendshape vectors as part of the model input enables a privacy-preserving approach, as these vectors abstract away identifiable facial features while retaining expression dynamics.

Second, by rendering 3D facial expressions from these vectors using Blender, researchers can generate large-scale synthetic datasets without requiring exhaustive data collection or manual annotation. This synthetic data generation pipeline enhances scalability and consistency across datasets.

In practical terms, the framework can be applied to emotion-aware VR/AR learning platforms, where detecting learners' affective states may guide content adaptation. In mental health applications, virtual assistants or therapeutic agents could respond empathetically based on users' expressions. In entertainment and gaming, emotional responsiveness may enhance player immersion and interaction. Furthermore, digital avatars in metaverse platforms can benefit from emotional expressivity and responsiveness, enabled by this architecture.

IV. Future Directions and Research Challenges

Although the framework demonstrates a promising design, future work is essential to validate its effectiveness. The construction of a labeled dataset linking blendshape vectors and rendered images to ground-truth emotion labels will be a necessary step. This dataset would allow for comparative experiments to assess the utility of each modality (image, vector, and their combination) on classification performance.

To handle dynamic expressions and temporal variation, future models may incorporate sequence modeling components such as LSTM networks or Transformer-based encoders. These temporal models could capture transitions between emotions and improve recognition of complex affective states. Cross-subject generalization will also be an important aspect to explore, ensuring the system performs robustly across individuals with different facial structures or expression patterns.

Additionally, there is potential to investigate zero-shot or few-shot learning approaches for recognizing previously unseen emotions or user-specific expressions. Overall, these future directions aim to improve the accuracy, scalability, and practical deployment of the proposed framework in real-world VR/AR systems.

V. Conclusion

We presented a novel framework for emotion recognition that utilizes blendshape data from Meta Quest Pro to generate 3D facial renderings and enable multimodal emotion classification. By synthesizing controllable, privacy-preserving data and combining visual and geometric modalities, the system offers a promising direction for affective computing in immersive environments. Our work lays the groundwork for future research on high-fidelity, real-time, and ethical emotion recognition systems that can operate in VR/AR contexts.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176). This work was supported by the Soonchunhyang University Research Fund (No. 2025-0011).

References

[1] Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on affective computing, 1(1), 18-37.

[2] Lou, J., Wang, Y., Nduka, C., Hamedi, M., Mavridou, I., Wang, F. Y., & Yu, H. (2019). Realistic facial expression reconstruction for VR HMD users. IEEE Transactions on Multimedia, 22(3), 730-743.

[3] Wang, S., Cheng, Z., Deng, X., Chang, L., Duan, F., & Lu, K. (2020). Leveraging 3D blendshape for facial expression recognition using CNN. Sci. China Inf. Sci., 63(2), 120114.

[4] Kuang, C., Cui, Z., Kephart, J. O., & Ji, Q. (2022, October). Au-aware 3d face reconstruction through personalized au-specific blendshape learning. In European Conference on Computer Vision (pp. 1-18). Cham: Springer Nature Switzerland.

[5] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

[6] Chaudhuri, B., Vesdapunt, N., Shapiro, L., & Wang, B. (2020). Personalized face modeling for improved face reconstruction and motion retargeting. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16 (pp. 142-160). Springer International Publishing.

[7] Lee, J. P., Jang, H., Jang, Y., Song, H., Lee, S., Lee, P. S., & Kim, J. (2024). Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface. Nature Communications, 15(1), 530.