Recognizing Challenging Behavior for Multiple Children with Intellectual and Developmental Disabilities Using the 3SLC Method

Jonguk Jung
Department of Computer and Information Engineering
Daegu University
Gyeonsan-si, Repulic of Korea
uwol6621@gmail.com

Yoosoo Oh
School of Computer and Information Engineering
Daegu University
Gyeongsan-si. Republic of Korea
yoosoo.oh@daegu.ac.kr

Abstract*— Challenging behaviors in children with IDD(Intellectual and Developmental Disabilities) can threaten their safety and the safety of others as well as the surrounding environment. Recording challenging behaviors in IDD children is currently done manually, which may lead to errors. We generate videos for each person in a multi-user setting and extract joint coordinate values to create composite joint information. In this paper, we propose the 3SLC method that combines Self-Attention, Long-Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN). The proposed 3SLC method analyzes the relationships between the joint composite information, extracts temporal features from the feature map, and analyzes spatial characteristics to predict the behaviors of children with IDD.

Keywords*— *3SLC, Joint Complex Information, IDD, Challenging Behavior, Pose Estimation

I. INTRODUCTION

IDD(Interactive and Developmental Disabilities) is a disorder that shows an overall disability to cognitive, social, and physical development. Children with IDD have difficulty overcoming various developmental tasks. Children with IDD have difficulty in learning, communication, and adaptation to daily life, and there are cases where they show challenging behavior in the process of experiencing difficulties [1,2]. Challenging behavior consists of self-harm, aggressive behavior, destructive behavior, and social maladjustment behavior [2]. Challenging behavior is a severe problem that can threaten the safety and environment of people around, not just the child. In particular, frequent aggressive behavior can cause physical and emotional damage to teachers or peers. Challenging behavior can cause educational and social exclusion of children with IDD.

Currently, most of the processes of recording and analyzing the challenging behaviors of children with IDD are manually conducted by managers such as teachers and therapists. Managers collect data by directly observing and recording the situation and frequency of actions or periodically checking behavior patterns [3]. However, handwriting tends to depend on the subjective judgment of the recorder and needs more accuracy, consistency, and objectivity because monitoring for a long time is limited [3]. In an environment where multiple children with IDD must be managed simultaneously through handwriting, there is a limitation that it is not possible to record the challenging behaviors of all children accurately. Direct observation and handwriting by managers can lead to data omission or distortion.

Therefore, we propose a method for automatically classifying the challenging behavior of children with IDD through 3SLC (3SLC) models (3Self-attention [4, LSTM[5], CNN[6]) to overcome the limitations of existing methods of handwriting the challenging behavior of IDD children. The proposed method generates 35 complex joint information consisting of 15 joint coordinate (x, y, z) information, 8 joint angle information, and 12 joint distance information in human object-specific images. We train a 3SLC model in which joint complex information is placed with three Self-attention and LSTM in parallel, and CNN is combined. The proposed method predicts children's behavior with IDD by analyzing the association between joint complex information through the 3SLC model and spatial characteristics from the feature map extracted with temporal features.

II. METHOD FOR RECOGNIZING INDIVIDUALIZED CHALLENGING BEHAVIORS IN MULTIPLE CHILDREN WITH IDD

A. Generation of joint complex information for Improved Recognition Accuracy

In our study, we focus on extracting only relevant joint coordinates from the joint data of actions performed by children with intellectual and developmental disabilities (IDD). Specifically, we exclude 18 joints that are either unnecessary for behavior analysis or too small to be reliably observed in

videos (left eye inner, left eye, left eye outer, right eye inner, right eye, right eye outer, left mouth, right mouth, left pinky, right pinky, left index, right index, left thumb, right thumb, left heel, right heel, left foot index, right foot index). After excluding 18 joints, we are left with a total of 15 key joints (nose, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle), which are used for training and predicting the behavior recognition model for IDD children.

Fig. 1. Figure 1 (a) shows 33 joints information from BlazePose, (b) shows 35 pieces of joint complex information from propoesd method.

This study calculates joint angles and distances from fixed joint coordinates, allowing for analyzing relative values in behavior learning and prediction. Precisely, 8 joint angles are calculated and extracted using key-joint coordinates: right elbow, left elbow, right knee, left knee, right armpit, left armpit, right hip, and left hip. In addition, twelve joint distances are computed from the joint coordinates, covering distances such as nose to left wrist, nose to right wrist, left wrist to right wrist, left ankle to right ankle, nose to left elbow, nose to right elbow, left wrist to left knee, right wrist to right knee, left wrist to right shoulder, right wrist to left shoulder, left shoulder to left knee, and right shoulder to right knee. The proposed method utilizes 35 features to analyze and predict challenging behaviors in children with intellectual and developmental disabilities (IDD), including 15 joint coordinates (x, y, z), 8 joint angles, and 12 joint distances. Figure 1 (a) shows 33 joints information from BlazePose. Mediapipe's BlazePose [7] is a model capable of inferring 33 joints, comprising 11 facial joints, 8 hand joints, and 14 body joints. Figure 1 (b) shows 35 pieces of joint complex information from the proposed method.

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5C2A07091326).

B. Structure of Method Recognizing Individualized Challenging Behaviors in Multiple Children with IDD

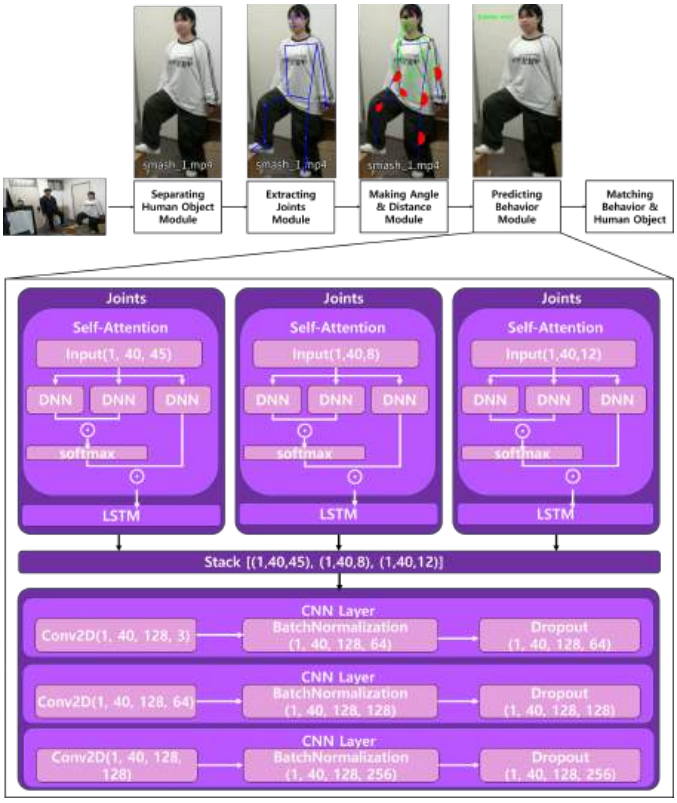

Fig. 2. architecture of the proposed multi-user recognition method for identifying challenging behaviors in children with IDD.

The proposed method takes video input containing human subjects. The Separating Human Object Module uses a pre-trained YOLOv8n segmentation model [8] to track human objects and saves individual videos for each detected object. The Extracting Joints Module applies BlazePose from Mediapipe to extract the x, y, and z coordinates of 33 joints from the video of a single human object. The Making Angle & Distance Module calculates joint angles and the distances between joints using the x, y, and z coordinates of 15 joints, generating complex joint information. The Predicting Behavior Module receives each human object's joint coordinates, angles, and distances and analyzes their relationships through a Self-Attention layer. This module compiles information from 40 frames to generate a heatmap that captures spatial and temporal features. It then utilizes CNN layers to analyze these features and classify behaviors. Finally, the Matching Behavior & Human Object Module compiles all predicted behaviors for each object in the video and saves them with object IDs in a single text file. Figure 2 shows the architecture of the proposed multi-user recognition method for identifying challenging behaviors in children with IDD.

III. EXPERIMENT

A. Construction of a video dataset consisting of 19 distinct actions

Fig. 3. Example of Challenging Behavior Video Dataset

To train the model for challenging behaviors, we constructed a video dataset. Since recording the challenging behaviors of children with intellectual and developmental disabilities (IDD) could violate human rights, we instead filmed 80 undergraduate students majoring in Special Education, trained to understand and simulate these behaviors, at four frames per second for 10-second clips. The dataset includes 19 types of behaviors:

- Five self-harming behaviors (self-hitting, self-biting, self-chewing, self-bumping, self-banging)
- 11 aggressive behaviors (hitting, kicking, extorting, pinching, scratching, striking, throwing, swinging, pulling, biting, choking)
- Two destructive behaviors (smashing, tearing)
- One no-action behavior

A total of 19 types of videos were filmed to create the dataset. During data preprocessing, we removed samples where joint positions were not visible, videos with fewer than 20 frames, and videos in which individuals other than the person acting were also captured. For videos with frame counts between 20 and 40, we repeated the last action until the frame count reached 40, ensuring the final dataset consisted of sequences with 40-time steps. Table 1 shows the number of samples per class in the constructed dataset containing 14,010 samples. Figure 3 shows an Example of a Challenging Behavior Video Dataset.

TABLE I. NUMBER OF SAMPLES IN THE PREPROCESSED DATASET

Behavior	Number of Samples per Class	
	Number of Frames before preprocessing	*Number of Samples After Preprocessing*
tear	71,592	798
smash	70,084	785
swing	36,204	784
kick	36,032	783
throw	35,854	777
scratch	35,781	776
self_bumping	35,770	803
hit	35,255	766
extort	35,226	762
pinch	35,072	760
choking	34,989	767
self_hitting	33,872	737
strike	33,690	750
pull	33,538	725
bite	32,960	717
no_action	32,943	713
self_bitting	32,671	711
self_chewing	30,808	680
self_banging	16,283	416

B. Performance Analysis of the Model Trained with Joint Coordinate Information

To train the model with the constructed dataset, we used the 99 coordinate values (x, y, z) for 33 joints of the human object in the video data. We trained and compared the performance of three models: an LSTM model consisting solely of LSTM layers, the LAC model consisting of LSTM, Attention, and CNN layers, and the MCD model consisting of MultiHeadAttention, CNN, and DNN layers. Table 2 shows the model evaluation with joint data.

TABLE II. MODEL EVALUATION WITH JOINT DATA

Model	Model Information		
	Model Constructure	*Train Accuracy*	*Validation Accuracy*
LSTM	Only LSTM	0.9725	0.7121
LAC	LSTM + Attention + CNN	0.6937	0.5023
MCD	Multihead Attention + CNN + DNN	0.9633	0.7217

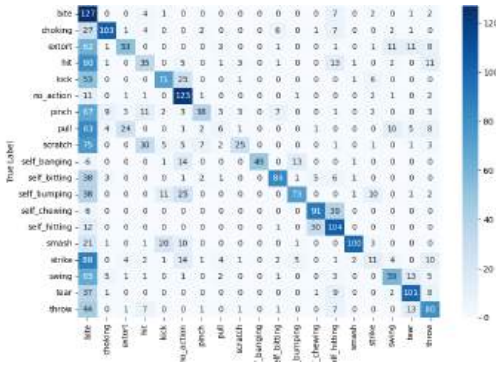

Fig. 4. Confusion matrix of MCD model

We found that the MCD model, which learned joint information about the human object, achieved the highest performance among the three models. However, we also observed that it was biased towards the first class, "bite," in the confusion matrix. Figure 4 shows the Confusion matrix of the MCD model.

C. Performance Analysis of the Model Trained with Joint Coordinates and Joint Angles

To solve the bias, we generated 8 joint angle values from the joint coordinate information extracted from the image dataset and used them for training. The data used for training consisted of 99 x, y, and z coordinate values for 33 joints, along with 8 joint angles for the right elbow, left elbow, right knee, left knee, right armpit, left armpit, right pelvis, and left pelvis, totaling 107 features. We trained the MCD model, composed of Multihead Attention, CNN, and DNN layers, to learn from joint coordinates and angles. We also trained the 2MCD model, which processes joint coordinates and angles in parallel using two Multihead Attention layers, CNN and DNN. Finally, we trained the 2SLC model, which parallelizes joint coordinates and joint angles through a Self-Attention layer and consists of LSTM and CNN layers.

TABLE III. MODEL EVALUATION WITH JOINTS AND ANGLES DATA

Model	**Model Information**		
	Model Constructure	***Train Accuracy***	***Validation Accuracy***
MCD	Multihead Attention + CNN + DNN	0.9770	0.7312
2MCD	2Multihead Attention + CNN + DNN	0.9631	0.7868
2SLC	2Self-Attention+LSTM+CNN	0.7639	0.7024

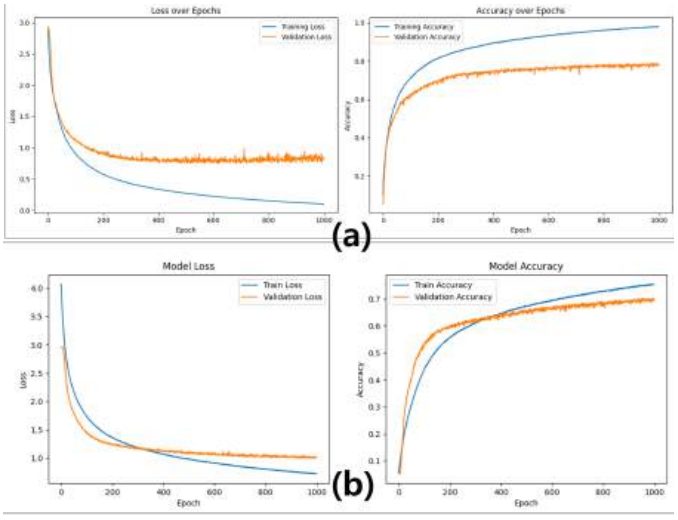

Fig. 5. (a) Accuracy and Loss Graph of the 2MCD model, (b) Accuracy and Loss Graph of the 2SLC model

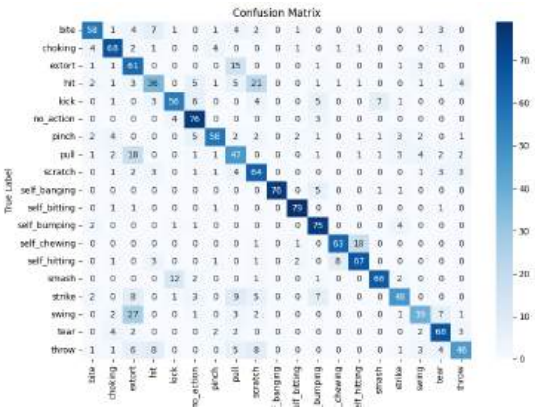

Fig. 6. Confusion Matrix of 2SLC model

Among the models trained with 99 joint coordinates and 8 joint angles, the 2MCD model achieved the highest performance. However, in the graph of the 2SLC model, the difference between the Training Loss and Validation Loss was insignificant, and the bias towards the first "bite" class was resolved. Therefore, we adopted a method that uses parallelized Self-Attention layers to analyze the correlations between joint coordinates and joint angles, followed by LSTM layers to analyze data for 40 frames. After generating a heatmap, the data is analyzed using CNN. Table 3 shows Model Evaluation with joints and angles data. Figure 5 (a) shows Accuracy and Loss Graph of the 2MCD model, Figure 5 (b) shows Accuracy and Loss Graph of the 2SLC model. Figure 6 shows the confusion matrix of the 2SLC model.

D. Performance Analysis of the Model Trained with Joint Complex Information dataset

To improve the performance of the action recognition model, we identified similar behaviors among the 19 classes. We merged the following actions into the same class: "pull" and "extort" for pulling actions; "hit," "scratch," "pinch," "throw," and "tear" for arm-swinging attacks; "kick" and "smash" for leg movements causing harm; "self-chewing" and "self-hitting" for self-injury to the head; and "strike" and "self-bumping" for body-related collision actions. As a result, the number of classes was reduced to 11.

TABLE IV. MODEL EVALUATION WITH JOINT, ANGLE AND DISTANCE DATA

Model	**Model Information**			
	Model Constructure	***class***	***Train Accuracy***	***Validation Accuracy***
3SLC	3Multihead Attention + CNN + DNN	19	0.9812	0.8033
3SLC with 11class	3Multihead Attention + CNN + DNN	11	0.9841	0.8593
SLC	Self-Attention+LSTM+CNN	11	0.9838	0.8527
2SLC	2Self-Attention+LSTM+CNN	11	0.9177	0.6857
3SLCD	3Self-Attention+LSTM+CNN +DNN	11	0.9755	0.8209

In this study, we aimed to enhance the performance of the human activity classifier by extracting 12 joint distance features from the joint coordinates obtained from video data. These features include the distances between the following pairs of joints: nose to left wrist, nose to right wrist, left wrist to right wrist, left ankle to right ankle, nose to left elbow, nose to right elbow, left wrist to left knee, right wrist to right knee, left wrist to right shoulder, right wrist to left shoulder, left shoulder to left knee, and right shoulder to right knee. We removed 18 joints that were either not used in the actions or were too small to be observed in the videos (left eye inner, left eye, left eye outer, right eye inner, right eye, right eye outer, left mouth, right mouth, left pinky, right pinky, left index, right index, left thumb, right thumb, left heel, right heel, left foot index, right foot index). Using 45 joint coordinates, 8 joint angles, and 12 joint distance features, which form the joint complex information, we trained five models: the 3SLC model (for classification into 19 classes), the 3SLC with 11 classes model, the SLC model (which learns 65 features with a single self-attention layer), the 2SLC model (which learns only joint angles and joint distances), and the 3SLCD model (which combines the 3SLC with DNN layers). The 3SLC with 11 classes model showed the highest performance. Therefore, we selected the 3SLC with 11 classes model to recognize challenging behaviors in children with IDD. Table 4 shows the evaluation with joint, angle, and distance data.

IV. Conclusion

In this study, we generated videos for each human object to recognize challenging behaviors in children with IDD. From the generated videos of each human object, we extracted joint coordinate information and computed joint angles and joint distances to create composite joint features. Additionally, we developed the 3SLC model using the generated composite joint information to identify challenging behaviors in children with IDD. To assess the performance of the 3SLC model, we conducted experiments using three different datasets: one containing joint coordinate data, another combining joint coordinate data with joint angle data, and a third including joint coordinate data, joint angle data, and joint distance data. As a result, we showed that the proposed 3SLC model achieved the highest performance.

In future work, we plan to design a model capable of learning by dividing the human body into different parts to improve the accuracy of challenging behavior recognition in children with IDD. By predicting behavior based on body part segmentation, we aim to enhance the accuracy of behavior prediction when a human object is occluded or when two behaviors are performed simultaneously.

Acknowledgment

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5C2A07091326).

References

[1] Bilias-Lolis. E, and Martin. A. M. “Challenging Behavior in Students with an Intellectual Disability: Promoting Understanding and Compassionate Educational Practice.” Journal of the American Academy of Special Education Professionals, vol.100, pp.111, 2019.

[2] Tevis. C, and Matson. JL. “Challenging behaviour in children with developmental disabilities: an overview of behavioural assessment and treatment methods.” *BJPsych Advances*, vol.28, no.6, pp.401-409, 2022

[3] Reyes-Martín. J, Simó-Pinatella. D, and Font-Roura. J. “Assessment of Challenging Behavior Exhibited by People with Intellectual and Developmental Disabilities: A Systematic Review.” *International Journal of Environmental Research and Public Health*, vol.19, no.14 pp.8701, 2022.

[4] Vaswani. A, “Attention is all you need.” *Advances in Neural Information Processing Systems*, 2017.

[5] Yu. Y, Si. X, Hu. C, and Zhang. J, “A review of recurrent neural networks: LSTM cells and network architectures.” *Neural computation*, vol.*31. no.*7, pp.1235-1270, 2019.

[6] Alzubaidi. L, Zhang. J, Humaidi. A. J, Al-Dujaili. A, Duan. Y, Al-Shamma. O, and Farhan. L, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions.” *Journal of big Data*, vol.*8*, pp.1-74, 2021.

[7] Bazarevsky. V, “BlazePose: On-device Real-time Body Pose tracking.”, *arXiv preprint arXiv:2006.10204*., 2020.

[8] Yue. X, Qi. K, Na. X, Zhang. Y, Liu. Y, and Liu. C, “Improved YOLOv8-Seg network for instance segmentation of healthy and diseased tomato plants in the growth stage.” *Agriculture*, vol.*13, no.*8, pp.1643, 2023.

[9] Doody. O, “Ethical challenges in intellectual disability research.” *Mathews Journal of Nursing and Health Care*, vol.*1,* no.1, pp.1-11, 2018.

Authors’ background

Your Name	Title*	Research Field	Personal website
Jonguk Jung	Master student	Machine Learning, HCI	https://fic-lab.com
Yoosoo Oh	Full professor	Machine Learning, HCI	https://fic-lab.com