Doctoral Thesis

Quantitative Diagnostic Support Method for Developmental Disorder Symptoms in Children

発達障がい児のための定量的な診断支援法

Prasetia Utama Putra

Department of Physics, Electrical and Computer Engineering,
Graduate School of Engineering,
Yokohama National University

Advisory Professor Keisuke SHIMA

January 2022

Abstract

Despite their high prevalence, many people often misinterpret symptoms of Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactive Disorder (ADHD) as willful or misconduct behavior. As conventional method is time-consumptive and susceptible to human-decision-making bias, a diagnostic support system to identify developmental disorder symptoms becomes requisite.

Previous studies have been proposing diagnostic support systems employing either bio-signal or behavioral test. Using bio-signals, previous works measured children's brain activity when they performed certain tasks. Then, using features extracted from the signals, the proposed approaches differentiated typical from disorder children with machine learning algorithm. As these approached required attaching sensors to children's body that might irritate children, other studies proposed to examine children's behavior directly using visual sensors or eye trackers.

This paper presents a diagnostic support system employing multiple visual sensors and eye tracker to measure children's behavior during indoor physical activity and the Go/NoGo task. The proposed system comprised group-level monitoring and individual-level monitoring. Group-level monitoring measured children's interaction with their environment and peers during playing activity in nursery schools. While individual-level monitoring examined children's game performance and gaze behavior in playing a game version of the Go/NoGo task. The proposed system used deep distance learning (DDL) to identify developmental disorder symptoms in children. It allowed the proposed model to measure similarity of a query to typical and disorder groups. Using DDL also enabled retrieval that provided evidence-based results. Estimation results were interpreted by employing SHAP values to provide specific information for the psychiatrist to identify developmental disorder in children.

This study includes four sections. First, we explained our study in estimating human activity with multiple cameras using Deep Neural Network model. Next,

based on the finding of the first study, we proposed behavioral monitoring system employing multiple Kinect sensors and RGB cameras. Third, we examined the relation between children's response and gaze behavior and investigated features relating to ASD and ADHD symptoms when they played Go/NoGo game. Last, we proposed a diagnostic support system employing Cluster Hard Triplet Loss to compute similarity between a query and typical and disorder groups and perform retrieval based using the query. It also provided an interpretation of the similarity score based on SHAP values.

Acknowledgements

I am sincerely grateful for all support and advice that my supervisor, Prof. Keisuke Shima, has extended to me during my study at Yokohama National University for six years. I deeply appreciate the opportunities that he gave me to grow personally and professionally in his laboratory. Though my research did not always go as I planned, he was very supportive and patiently counselled me to learn from the failures and mistakes that I made. He was also keen to give career advice that helped me to plan my career in academia.

I also want to express my sincere gratitude to Prof. Hamagami, Prof. Ochiai, Prof. Ichige, Prof. Sugimoto, and Prof. Otsuka for their valuable comments and feedbacks that enriched my dissertation. To all members of Shima laboratory, I would like to thank them for their supports and hospitality.

I want to show my appreciation to my collaborators, Prof. Shimatani and Prof. Sergio, who assisted me in conducting my research. I felt so fortunate to have a chance working with you. To my parents, sisters, and friends who supported me during my graduate school, I want to say that I will always be in debt to you for emotional support that you provided to me.

Lastly, I want to thank my best friend, Hiro, who has been keen and patient to hear my troubles in conducting my research. This study could not be realized without findings of my collage researchers; I am obliged to express my appreciation to all of you that enrich science.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Purpose

People often misinterpret invisible disorders symptoms such as impulsivity and hyperactivity as willful misconduct and poor character. Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactive Disorder (ADHD) are examples of invisible disorders.

In Japan, the prevalence of Autism Spectrum Disorder (ASD) symptoms among children has been estimated to be between 1.9% and 9.3% based on parent and teacher reports [5]; in the USA, about 1 of 54 children was diagnosed with ASD in 2020 [6], while in 2016, a study found that 9.41% of children had Attention Deficit Hyperactivity Disorder (ADHD) symptoms [7].

Conventional diagnostic methods involve monitoring children's behavior by psychiatrists with behavioral assessment checklist. The results of behavioral checklist, however, are prone to human decision-making bias and time-consumptive. Previous researches have attempted to resolve this issue by utilizing human biosignal such as EEG [8] and fMRI [9] to differentiate typical from disordered children. Thought previous findings suggested a promising performance of those methods in diagnosing the disorders, they neglected the urgency to develop a decision support system to identify high-risk disorder children. Also, methods using fMRI and EEG required experts' skills to be operated, hindering their

daily applications.

In contrast, psychiatry studies recognize disorder symptoms by employing behavioral tests that are simple to be carried out. The Go/NoGo task [10–12], and visual attention test [13] are example of behavioral test. Previous studies have discovered a significant difference between ASD and typical children during a Go/NoGo task. The task requires a subject to react to the Go stimulus and inhibit their reaction to the NoGo stimulus [12]. The stimuli can be represented by visual objects with different colors and shapes [10, 12] or by sounds with different frequencies [11]. Studies on ASD children's gaze behavior have observed that the ASD group was slower to adjust their gaze to the stimulus position during eye-tracking measurement of joint attention [13], and faced difficulty in modulating their gaze during face-to-face conversation [14].

Other approaches have proposed behavioral monitoring system that automatically tracked children's movement to understand the difference between typical and disorder children. The systems can be classified into a marker and marker-less methods.

Marker systems monitor children's behavior by attaching markers to their bodies. Three-dimensional information of an infant's joint can be obtained with a motion capture system that tracks the movement of markers on the subject's body [15]. Marker-less approaches use a visual sensor to eliminate the use of electrical markers in monitoring children's activity. Previous studies used a single RGB camera to measure infants' general movement [16] and to monitor toddlers' activity in a nursery room [17].

This study aims to diagnostic a decision support system employing behavioral test and monitoring system. The proposed method included group-level and individual-level monitoring system. Group-level involved marker-less behavioral monitoring system that tracked children's activity in a nursery room with multiple RGB-D sensors. Individual-level used a game version of the Go/NoGo task to identify children's impulsivity and inattentiveness by measuring their response and gaze behavior during the game. For that purpose, this paper discusses the

following topics:

- A deep neural network to estimate human activity

  Conventional CV methods have dominated multi-view human activity recognition system. Those methods involved sophisticated features extracting methods to combine features from multiple cameras. This study proposed a DNN model employing shared-weight to estimate human activity [1]. These results showed DNN based method could achieve competitive recognition rate using only 2D RGB inputs.

- Marker-less monitoring system to model behavior children

  Previous studies have found that ASD and ADHD children showed excessive physical movement, constantly changing activity, and little interest in peers. This paper proposed a marker-less monitoring system employing multiple RGB-D cameras [2]. We used OpenPTrack with Kinect sensors to track children's activity in the nursery room. Then, we model children's behavior with PetriNet model. The study represented typical and ASD subjects with four features extracted from the model and statistically analyzed the difference between the groups.

- Study on Response and Gaze Behavior during the Go/NoGo Task

  Findings of previous works on behavioral test of ASD/ADHD symptoms suggested a statistically significant difference between typical and their ASD peers in their response and gaze behavior [3]. This paper presents a serious game of the Go/NoGo task to measure subjects' response and gaze behavior during the task. We recruited 59 university students to take part in the experiment and analyzed the relationship between their gaze and response when they played the game. We performed statistical analysis and clustering to understand the patterns in their features. We also used our proposed game (CatChicken game) to measure response and gaze behavior of typical and ASD children. The study investigated whether features ex-

tracted from this information could identify the symptoms by conducting statistical analysis.

- Diagnostic support system with deep distance learning

  Findings of previous works on behavioral test of ASD/ADHD symptoms suggested a statistically significant difference between typical and their ASD peers in their response and gaze behavior [3]. This paper presents a serious game of the Go/NoGo task to measure subjects' response and gaze behavior during the task. We recruited 59 university students to take part in the experiment and analyzed the relationship between their gaze and response when they played the game. To understand the patterns in their features, we performed statistical analysis and clustering. We also used our proposed game (CatChicken game) to measure response and gaze behavior of typical and ASD children. The study investigated whether features extracted from this information could identify the symptoms by conducting statistical analysis.

1.2 Proposed Method

Our proposed method (Fig. 1.1) comprises group and individual levels monitoring. Group-level monitoring system aims to quantify children's playing behavior and identify high-risk disordered children. The system measures children's activity with multiple cameras and evaluates their interaction with the environment and their peers.

Contrary to group-level monitoring, individual-level monitoring system assess children's performance during specific task. This study uses a game version of the Go/NoGo task to measure children's response and gaze behavior. During the experiment, subjects respond to Go/NoGo stimulus by pressing a space bar on the keyboard. The system tracks their gaze movement with an eye-tracker attached on the monitor.

Finally, using the features from group and individual level monitoring, deep

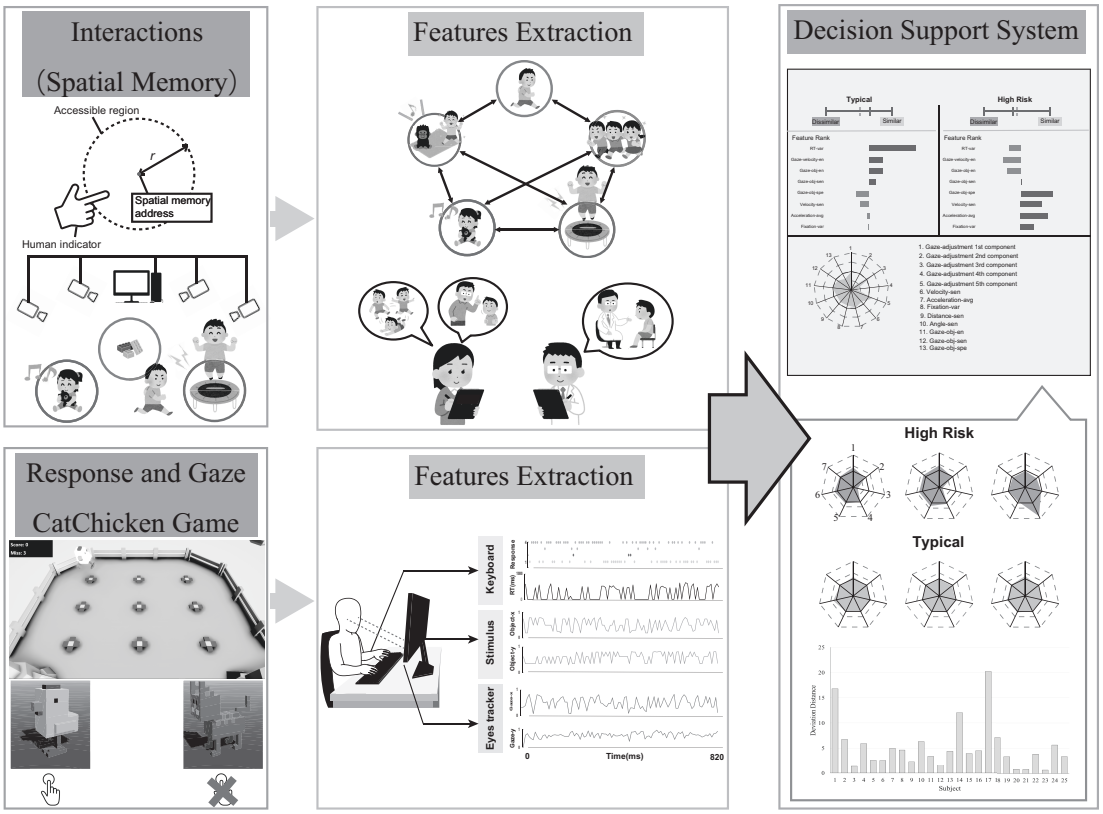


Fig. 1.1: A diagnostic support system to identify developmental disorder symptoms in children. The system comprises group and individual levels monitoring.

distance learning measures identify developmental disorder symptoms by measuring similarity between a query and the support vectors. Model agnostic model interprets the estimation results of the model, providing comprehensive information to the psychiatrists about the estimation results of individuals.

1.2.1 Group-Level: Children's Behavior Monitoring with Multiple Cameras

Group-level monitoring measures children's interaction with their environment and their peers utilizing multiple RGB-D cameras. Using RGB-D images, the proposed system tracks movements and positions of children and objects to estimate children's activity continuously. Next, the system quantifies their play-

ing behavior to provide comprehensive information for the psychiatrists, enabling them to make a better diagnosis.

1.2.2 Individual-Level: Children's Response and Gaze Modulation during the Go/NoGo Task

Individual-level monitoring offers a standardized test that is less susceptible to bias that presents in a natural playing environment. Individual-level monitoring evaluates children's behavior during the Go/NoGo task. The system measures children's game performance and gaze movement to identify irregularity in their response and gaze modulation. Spatial and gaze-adjustment features are extracted to represent the behavior of a child during the game.

1.2.3 Deep Distance Learning to Identify Developmental Disorder Symptoms in Children

To help psychiatrists make a better judgment in identifying developmental disorder symptoms, a diagnostic support system (DSS) must provide evidence based and interpretable results to the users [18]. Our proposed system utilizes deep distance learning and Shapely value to provide interpretable evidence-based results.

As symptoms of invisible disorders (e.g., ADHD and ASD) overlap among disorders, using DSS to determine the class of disorders may yield inaccurate results. Deep distance learning (DDL) allows the proposed system to perform similarity measurement between a query and support dataset, if a single query lies in the middle of several centroids of the support sets, then it will obtain equivalent similarity scores. DDL also enables the system to perform retrieval to fetch similar data to the query from the support set, which the psychiatrists can consider when making a diagnosis for patients.

The proposed system interprets the estimation results of DDS with SHAP values. The algorithm allows the proposed system to tell the users which feature-

related symptoms are pronounced in the patients and why the DDL produces such estimation results. The interpretable results help provide comprehensive metrics that psychiatrists can use to identify developmental disorder symptoms in children.

1.3 Dissertation Outline

This thesis presents a novel Diagnostic Support System that comprises group-level and individual-level monitoring systems (Fig. 1.2). Group-level system monitored children's playing behavior to identify hyperactivity and impulsivity symptoms. Individual-level system measured children's response and gaze behavior to measure their inattentiveness and impulsivity symptoms.

Chapter 2 discusses the previous works related to this study. It reports the benefit of using marker-less over marker monitoring system to monitor children's behavior. This chapter also includes the reason this study developed a serious game of the Go/NoGo task to measure children's motor response and gaze behavior. Last, it reviews the concept of diagnostic support system and how it can help psychiatrists in making better judgment to identify developmental disorder symptoms.

This thesis explains the detail of the proposed model in Chapter 1 that includes group-level and individual-level monitoring system. Chapter 3 and 4 describes our study related to group-level monitoring. We investigated on how we could estimate human activity using RGB images and DNNs model in Chapter 3. While in Chapter 4, we reported our behavioral monitoring system using OpenPTrack employing multiple Kinect sensors.

Chapter 5 reports an investigation of subjects' response and gaze behavior. We developed a serious game version of the Go/NoGo task to measure their response and gaze behavior when they played the game. The experimental results suggested a significant positive relationship between the participant's response and gaze trajectory area. Also, this chapter presents the results in ASD chil-

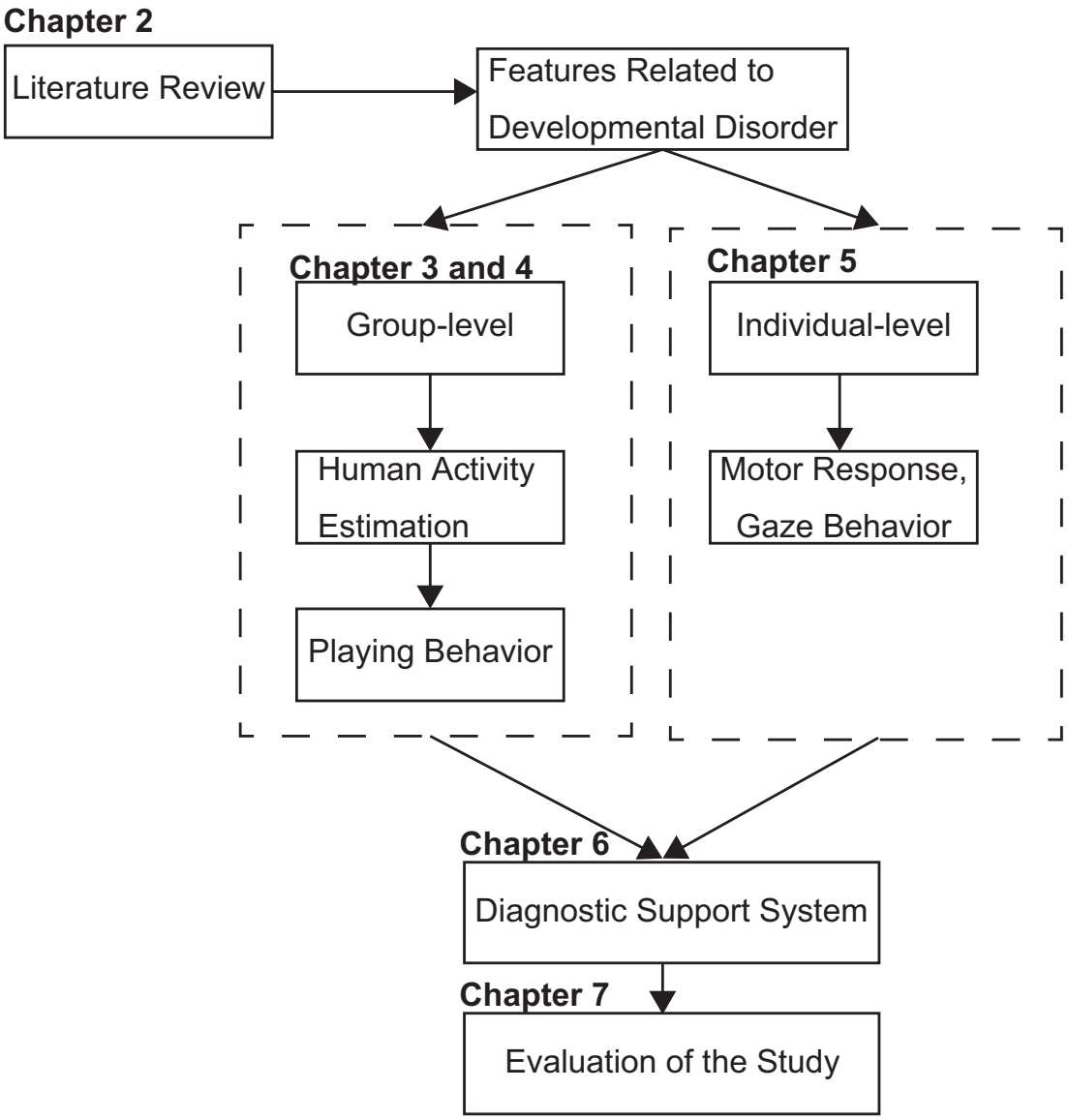


Fig. 1.2: Flowchart depicting the structure of this dissertation.

dren's response and gaze behavior when they played the CatChicken game. The statistical comparison showed a significant difference in gaze modulation behavior between typical and their peers with ASD children.

Study of diagnostic support system to identify ASD symptoms is presented in Chapter 6. The proposed system employed Deep Distance Learning (DDL) and SHAP value to provide interpretable evidence-based results. The DDL enabled

the system to produce evidence-based results and learn new class with only few samples. And the SHAP value allowed the proposed DSS to interpret DDL's similarity score.

Last, we discussed the results, limitation, and future direction of our study in Chapter 7.

Chapter 2

Related Work

2.1 Children's Behavior Monitoring with Marker and Marker-less Methods

Marker system used sensors or electrical markers attached to children's bodies to monitor their activity continuously. Using a motion capture system, Meinecke *et.al* [15] tracked infants' body movement by placing 18 markers placed on their bodies. They modeled the body movement using a rigid biomechanical model that was represented by 53 parameters. The experimental results suggested that the movement complexity of typical and pathological children differed, in which the latter group's movements were more monotone than the former. Although the proposed system achieved promising results, the equipment's setup process was complicated and required many markers attached to the infants. To tackle that problem, instead of a motion capture system, [19] utilized an accelerometer to monitor infants' body movement. Four accelerometer sensors were attached to the subjects' hands and feet and their spontaneous movement was monitored for 20 minutes. Although the biomechanical model used in this study had fewer parameters, classification results showed that the proposed approach outperformed the accuracy of [19].

Attaching marker to children's body may harm them or influence their behav-

ior during the experiment. Marker-less methods utilized visual, lidar, or radar sensors to estimate positions of subjects and objects. Previous studies utilized a single RGB camera to measure infants' general movement [16] and to monitor toddlers' activity in a nursery room [17]. While the former monitored individual infant's behavior continuously, the latter faced occlusion problems when monitoring a group of children's behavior.

2.2 Children's Behavior During the Go/NoGo Task

Go/NoGo task requires a subject to respond to the Go stimulus, but inhibits his action to the NoGo stimulus. This task can be classified as an action restraint since if the NoGo stimulus appears, then the subject must inhibit his action before execution [12].

Previous researches [10, 12, 20] represented the stimulus on a monitor as two distinct characters, colors, and shapes. Some of them also used sound with a different frequency to show when the participant had to respond and when they did not have to [11].

During the Go/NoGo task, a participant can respond by pressing a spacebar, clicking a left mouse, or touching a screen [20]. Response time (RT) defines the time difference between when the stimulus appears and when the participant reacts. Standard deviation, coefficient variability, and kernel density estimations were used to calculate its variability (RT Var) [20].

Since it was not invasive, many previous works used the Go/NoGo task to identify developmental disorder symptoms.

Bezdjian *et.al* [10] found that ADHD children showed higher NoGo errors and higher reaction time variability. The study also stated that Go error related to both verbal and IQ performance of children, while Go RT only correlated to the latter. Another work furthered this idea by conducting a meta-analysis of RT-var in ADHD. The results showed RT-var was specific not only to ADHD

symptoms but also to general mental disorders correlating with working memory and behavior inhibition.

2.3 Gaze Modulation of ASD CHildren

Since impairment in basal ganglia affects a person's visual attention, previous studies investigated the difference between typical and disordered children by measuring their gaze behavior [13, 21]. They tracked the subjects' eye movement and analyzed the gaze modulation when the subjects performed visual attention tasks [13] or interacted with others [14].

Difficulties in modulating gaze presented in children with Autism Spectrum Disorder (ASD) during fixation task experiments and when having dyadic interaction [13, 14]. ASD children showed lower gaze modulation and could not focus on their interlocutors' face during the interaction: they got distracted with the surroundings more often than the control group.

2.4 Application of Serious Game for Children with Developmental Disorder Symptoms

A serious game intends to bring an entertaining and non-invasive method to its users. They can engagingly learn or perform certain skills, e.g., visual attention, language, and social skills. In mental health, most previous works used serious games to rehabilitate disordered patients [22–24]. The system implementation involved the addition of game elements into the training task to enhance users' motivation.

The games required the patients to accomplish objectives relating to the disorder's symptoms. Games proposed by Faja *et.al.* [22] and Beaumont *et.al.* [23] trained ASD children's face recognition and social skills by challenging them to identify characters' face morphological features and to decode their thought and emotion. BrainGame Brian [24] delivered working memory, inhibition, and cog-

nitive task as a role-playing game, in which a player took part in those tests by solving problems of characters in the game.

2.5 Overview of a Diagnostic Support System

Clinical diagnostic support systems (CDSS) aims to help clinicians in making diagnosis by combining their knowledge with suggestion provided by the system [18]. The system comprises knowledge-based and non-knowledge-based systems. The latter uses expert medical knowledge, literature, and patient-directed information to create a rule-based decision support system. It works by retrieving data from patients and database to evaluate the rule and producing suggestions. Conversely, the former employs machine learning to leverage information of patients in making outputs.

The functions of CDSS are to improve patient safety [25], clinical management [26], cost containment [27], and diagnostic support [28]. Clinical diagnostic support (CDS) provides suggestions to clinicians in making diagnostic of patients. Previous works' findings employing non-knowledge-based diagnostic support showed high accuracy in detecting neuropathies, diabetic retinopathy, and tumor. In other studies, CDS was employed to enhance radiology images [29].

One of issues caused slow implementation of CDS were physicians' bias and negative perception because of uninterpretable suggestions. This paper attempted to resolve this issue by employing interpretable deep distance learning. Deep distance learning computed similarity score between a query and support sets and retrieved similar support sets from the database. SHAP value [30] interpreted the similarity score, providing interpretable suggestions.

Chapter 3

A Deep Neural Network Model for Multi-view Human Activity Recognition

3.1 Introduction

Occlusion often occurs when observing human activity using a single camera [31]. The occlusion causes information loss that leads to failure of activity recognition in single-view human activity recognition. Previous studies have attempted to resolve this issue with multiple-view technique that uses multiple cameras to compromise information loss when occlusion appears in a single camera [31, 32].

Multi-view human activity recognition (MVHAR) comprises conventional computer vision (CV) and DNN-based methods. Conventional CV methods have represented human movement as low-level features such as histogram of gradient (HoG) [33], silhouettes [34], and optical flow [35], extracted from the sequence of RGB images. Then, the features were categorized with classification algorithms or transformed to higher-level representation. Previous works estimated human activity either by combining features from multiple inputs followed by classifier

algorithm or by estimating actions from a single input followed by score fusion or vote. Studies employing conventional CV aimed to improve the recognition rate of MVHAR with effective features-extraction and features-utilization algorithms.

In contrast with conventional CV methods, DNN approach combines features extractor and classifier into single pipeline. A DNN model automatically discovers representation and recognizes the patterns in given data using an end-to-end learning algorithm [36].

Previous works on DNNs in MVHAR have involved early [1] or late fusion to combine multiple inputs [37]. Early fusion resulted in a DNN model with a modest number of parameters, which combined features from the early layer [1]. However, the combinations of inputs in an early layer with this approach may cause highly variant features, making it prone to over-fitting. Meanwhile, late fusion combined features from multiple cameras by treating inputs individually with multiple models [37,38]. Individual models in this approach may have fewer variant features but a high number of parameters, which consume more memory. Other DNN approaches have attempted to solve the multi-view human action recognition by employing multimodal inputs [39–43], multi-task training [38], and cross-view learning [44, 45] algorithm.

This paper presents a novel DNN model using a shared-weight application for multi-view human action recognition. The shared-weight application enabled the model to perform late fusion with fewer parameters than the multi-block technique. The model used multi-view images as inputs to produce multiple hypotheses. Then, using the score-fusion, the model computed the final prediction. Since the prior knowledge about informative inputs among multiple views was unknown, we trained the proposed model to treat each prediction score equally with arithmetic mean or weight the hypothesis with geometric mean. The model applied an attention network to filter out uninformative features from each view.

The model comprises pre-trained CNNs, attention layers, RNN, and Softmax layers. The study conducted exploration studies for structural optimization and performed transfer learning with pre-trained CNNs to prevent over-fitting in the

training process. We compared the performance of the proposed model performance to that of the-state-of-the-art application on IXMAS [46] and i3DPost [47]. We conducted an online evaluation and comparative study with the single-view model to determine its efficiency in the actual situation.

3.2 Proposed Model [1]

Our proposed model comprises CNNs, attention layer, RNN, and Softmax blocks (Fig 3.1). We extracted features from multiple-view inputs by feeding them to CNNs block. Then, attention-layer filtered out uninformative features by applying attention mask to the latent variables. The proposed model employed RNN to understand temporal information before computing probability of activity classes with Softmax layer. The proposed model shared the weight of pre-trained CNNs and RNN across multiple inputs but used different attention and Softmax layers for each input.

This study involved an examination of the pre-trained models VGG-19 and VGG-16 [48] comprising five blocks with different numbers of CNNs. This paper refers to I-th CNN in N-th block as `block`N`_conv`I.

3.2.1 Attention Layer

The proposed model assumed that significant transformation occurred only in certain parts of image sequences when subjects performed actions. Thus, it should focus on certain frames in estimating activities. To filter out uninformative features, our proposed model employed an attention layer [49] that weighted important features with higher probability and the others with lower probability.

Given the feature vector F of shape $T \times G$, the attention mask was computed by averaging attention scores over G. The first step to determine relevant features was to estimate attention probability at each time step for the G dimension. For the feature map at the t-th time step f_t, attention probability was given by

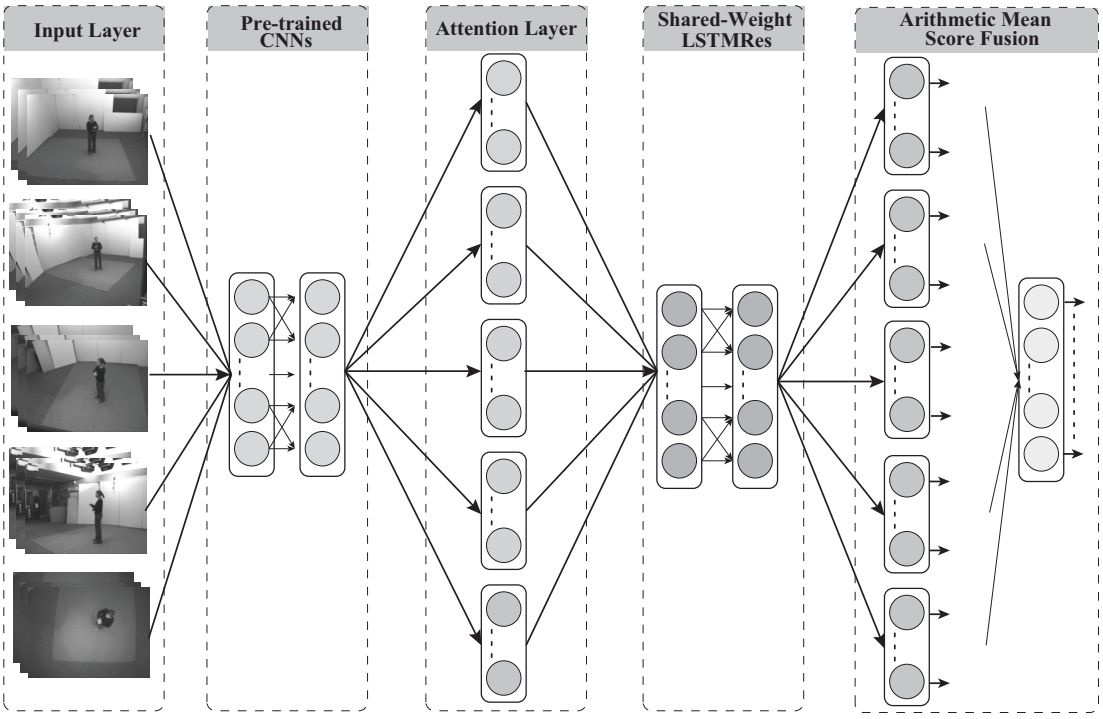


Fig. 3.1: Architecture of the proposed model. Pre-trained CNNs and LSTMRes were shared across inputs.

$$s_t = g_{\text{att}}(f_t; \theta_t) \tag{3–1}$$

$$\alpha_t = \text{softmax}(s_t) \tag{3–2}$$

where g_{att} was an attention network with weight θ_t, and s_t was the attention score map for the feature map. The attention score α_t was the probability produced from Softmax function incorporating the subject of interest with a higher probability than the rest. Dense, convolutional, and RNN layers can be used as attention networks [50]; the proposed model employed a dense layer for the attention network because, in the preliminary experiment, we found CNN and RNN caused over-fitting. Besides, the proposed model intended to filter out uninformative frames and not features in each frame, such as in sequence2sequence model [51].

After computing attention probability at each time step, the relevant features

were calculated using

$$\hat{f}_t = f_t \odot \alpha_t \tag{3–3}$$

where $\odot$ represents the element-wise operator or the Hadamard product [52]. Eq. 3–3 weights features extracted from pre-trained CNNs with α_t.

3.2.2 Residual Learning in LSTM

A long short-term memory (LSTM) architecture [53] was proposed to solve the problem of vanishing and exploding gradients associated with conventional recurrent neural networks (RNN) [36]. The architecture, however, still can suffer from degradation problems caused by deeper neural network structure [54]. Residual learning was proposed to tackle this issue by introducing a shortcut connection from the earlier to the later layers that helps the earlier layer get a-"fresh"-gradient from the latter one during backpropagation [55].

In contrast to the highway network approach [56], residual learning formulation [55] involved an identity shortcut to ensure ongoing learning. Residual function $H(z_i)$ could be expressed as:

$$H(z_i) = F_i(z_i, W_i) + W_s z_{i-m} \tag{3–4}$$

where $F(z_i, W_i)$ and z_{i-m} represent the original mapping and output from the earlier layer, respectively and W_s was a linear projection that was used when the dimension between $F(z, W_i)$ and z_{i-m} was unequal, as realized via linear mapping.

In LSTM, residual mapping could be accomplished by introducing a shortcut connection to the adjacent layer, from layer t to $t+1$ [57] (Eq. 3–5), or by establishing a connection to the memory cell [1] (Eq. 3–6, implementation: Fig 3.2).

$$h_t = o_t \odot tanh(C_t) + h_{t-1} \tag{3–5}$$

$$h_t = o_t \odot tanh(C_t + W_s x_t) \tag{3–6}$$

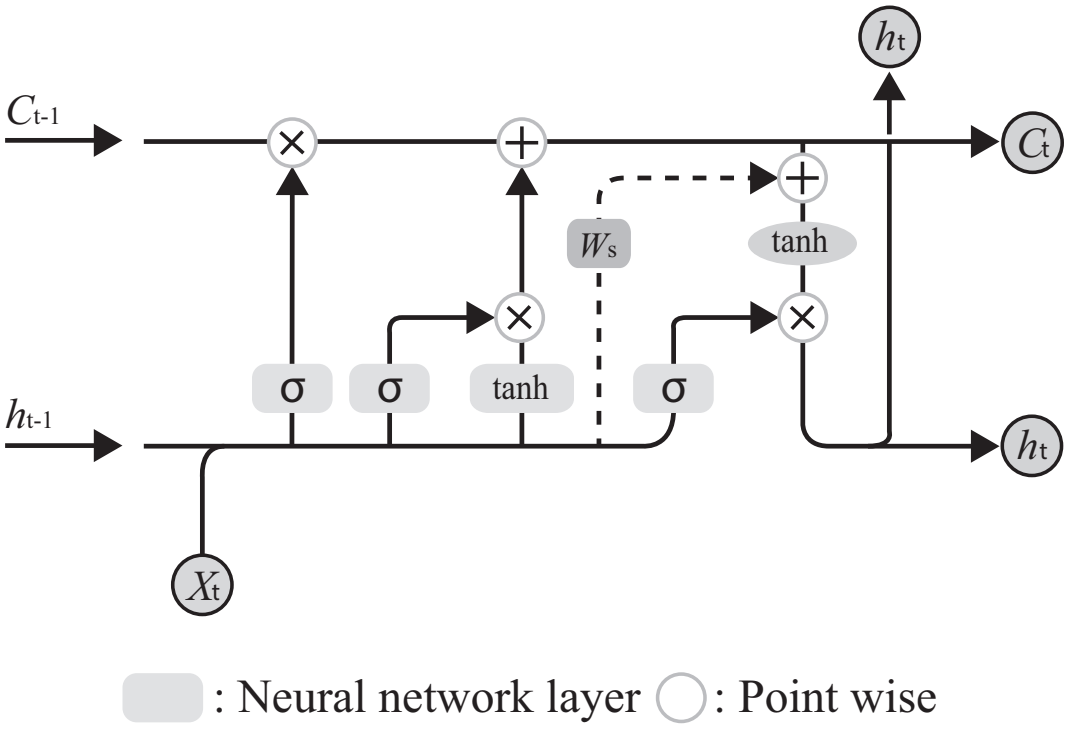


Fig. 3.2: Architecture of LSTM with residual learning. Implementation of residual learning in LSTM with shortcut connection after forgetting of old information and addition of the new information. The dotted line shows shortcut connection.

Here, o_t, C_t, h_t represent the output gate, memory cell, and hidden units, respectively.

3.2.3 Score Fusion

Arithmetic or geometric means are used to combine prediction scores from Softmax layers. With the former, scores from all cameras were treated as a mixture, while the latter allows one prediction result from a single camera to veto other outcomes. The proposed model calculated final prediction scores using arithmetic mean (Eq. 3–7)or geometric mean (Eq. 3–8).

Here, y_{ac} represent the probability score of an action a from camera c. M and N are respectively the total number of actions and cameras.

$$Y_a = \frac{\sum_c^N y_{ac}}{N} \tag{3–7}$$

$$Y_a = \frac{\sqrt[N]{\prod_c^N y_{ac}}}{\sum_a^M \sqrt[N]{\prod_c^N y_{ac}}} \tag{3–8}$$

3.3 Method

3.3.1 Datasets and Evaluation Metric

The IXMAS dataset [46] is a benchmark in MVHAR algorithm evaluation that comprises videos of 12 subjects performing 13 actions: watch checking, arms crossing, head-scratching, sitting, getting up, turning around, walking, waving, punching, kicking, pointing, picking something up, and throwing. Videos were recorded using five cameras at 23 fps. Subjects performed each action three times with free positioning and orientation.

The i3DPost dataset [47] was recorded using eight synchronized cameras with a resolution of 1920x1080 and 25Hz progressive scan. The eight subjects performed 12 actions (walking, running, jumping, bending, waving, jumping in place, sitting-standing, running-falling, walking-sitting, running-jumping-walking, hand-shaking, and pulling), creating 96 multi-view videos of human activity.

The proposed model's performance was evaluated with categorical cross-entropy loss, classification accuracy, and F1-score metrics. The accuracy rate was computed by averaging top-1 accuracy for given data, while F1-score was the average F1-score for all classes.

3.3.2 Pre-processing and Learning

To reduce distortion in images and ensure the features were on the same scale, RGB-normalization and feature standardization were performed in pre-processing.We computed individually mean and standard deviations for feature

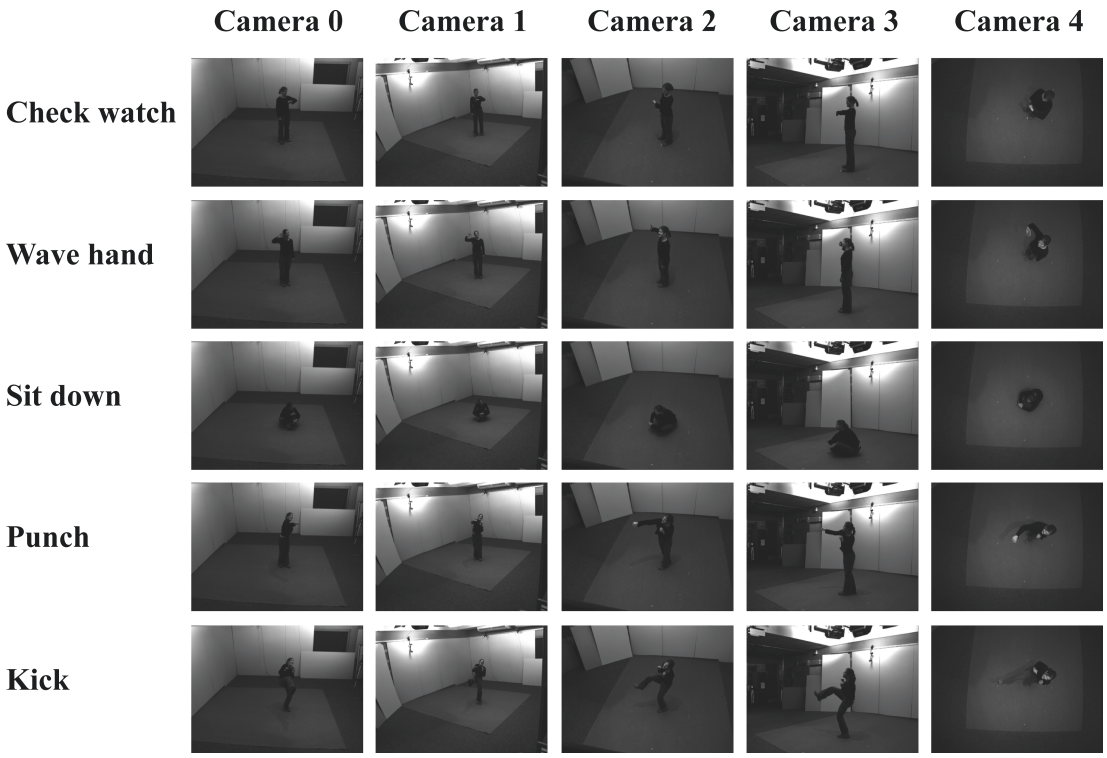


Fig. 3.3: Sample clips from IXMAS dataset. Five cameras were used to record activities of subjects.

standardization for each dataset. And gamma correction applied to images for experiments using the IXMAS dataset; the gamma value was 1.5.

In the experiments, the proposed model was trained with scenario I (LSTMRes evaluation), scenario II (evaluation of the pre-trained model, MSLTMRes and score fusion, and implementation of online classification), and scenario III (a comparison of the proposed model with state-of-the-art methods) (Table **??**). In all scenarios, backpropagation with RMSProp optimizer [58] was used.

3.4 Results

3.4.1 Exploration Studies

This section details the results of exploration studies using the IXMAS dataset. The experiments included:

1. a performance comparison of LSTMRes with LSTMResKim [57], LSTM, and Convolutional LSTM (ConvLSTM) [59],
2. an investigation on the impact of fine-tuning pre-trained CNNs with VGG-19 and VGG-16; other models such as ResNet [55] and Inception [60] were not used because they impaired the recognition rate of the proposed model,
3. an evaluation on using a multiple-block approach and shared-weight technique, and
4. a comparison of features fusion with score fusion using arithmetic and geometric means.

The experimental protocols used in the experiments varied (Section Pre-processing and Learning). And in every experiment, we used the most optimal structure for the succeeding experiment.

A LSTMRes vs LSTMResKim

Fig 3.4 depicts the performances of LSTM, LSTMRes, and LSTMResKim on IXMAS based on training and validation errors. The results suggested that validation errors of the proposed model with the LSTMRes were lower than that with LSTMResKim. The training error with LSTMRes decreased steadily throughout the learning process. But instability appeared in those of LSTMResKim after the 60th iteration.

In contrast, the LSTMRes approach exhibited slightly lower training and validation loss than LSTM and ConvLSTM. The performances of LSTM and LSTMRes were identical. These outcomes indicated that performing residual learning

in the LSTM memory cell provides insignificant improvement with the model. In consideration of these results, we used LSTMRes used for the rest of the experiments reported here.

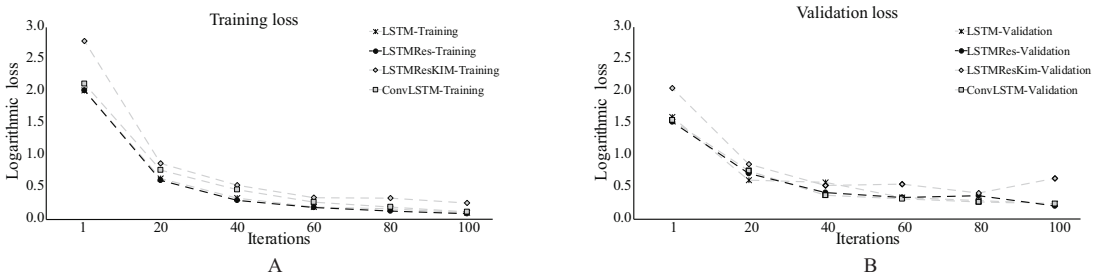


Fig. 3.4: Training and validation errors of LSTM, LSTMRes, LSTMResKim, and ConvLSTM. From the start until the end of learning, LSTMResKim had lower performance than the others.

B Pre-trained CNNs

VGG-16 and VGG-19 models trained with the ImageNet dataset were examined as CNN blocks for the proposed model. We conducted experiments using the intermediate `block4_pool` and the last `block5_pool` layers to find an appropriate pre-trained model and clarify the effect of fine-tuning. The first experiment used the intermediate layer as a feature extractor without fine-tuning the parameters, while the second applied it. The last experiment was conducted by fine-tuning the CNNs (from `block4_conv2` to `block5_conv3`).

Figure 3.5 illustrates the performance of the proposed model with different CNN blocks. The results showed three things. First, employing VGG-16 as a CNN block produced higher DNN model accuracy. The lowest recognition rate with VGG-19 in the intermediate layer experiment was higher than that with VGG-16. But the proposed model achieved an average increase in accuracy rate by $1.14 \pm 0.89\%$ using either the intermediate or last-layer of VGG-16. Second, fine-tuning the last block of pre-trained CNNs improved insignificantly ($n = 396$, average $p > 0.05$) the proposed model's performance. Fine-tuning VGG-16 and

VGG-19 improved respectively the accuracy rate by 4.29% and 2.52%. Third, fine-tuning of pre-trained CNN parameters may impair the performance of the proposed model, whose accuracy rate decreased by 2.02% with fine-tuning `block4` (intermediate)

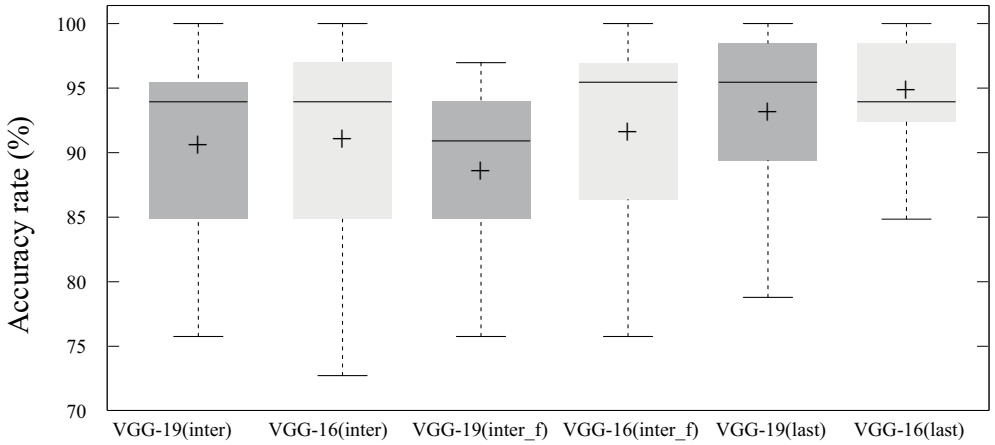


Fig. 3.5: Accuracy of the proposed model with different pre-trained CNNs. Average recognition rates of proposed model: VGG-19(intermediate): 90.40%; VGG-16(intermediate): 90.90%; VGG-19(fine-tuned intermediate): 88.38%; VGG-16(fine-tuned intermediate): 91.41%;VGG-19(last): 92.92%; VGG-16(last): 94.69%.

C Shared Weight, no MSLSTMRes

We previously found that MSLTMres yielded higher accuracy than the baseline model [1]. However, the recognition rate came at the expense of computational time and parameter numbers.

Given the benefits of shared-layer DNN in language modeling [61], we investigated related effects in multi-view action recognition, sharing the pre-trained VGG-16 and stacked LSTMRes of the proposed model across inputs from all cameras. Different attention layers were used for different views, and feature fusion was used to compute action probability.

An improvement ($n = 396$, $p = 0.617$) was observed with the use of shared-

weight LSTMRes in the proposed model (Fig 3.6). The proposed model exhibited a 1.01 % higher accuracy than with the use of MSLSTMRes. Shared-weight application also resulted in fewer parameters (the proposed model: 70,304,393, [1]: 351,323,711) and lower complexity with the proposed model, improving computation time.

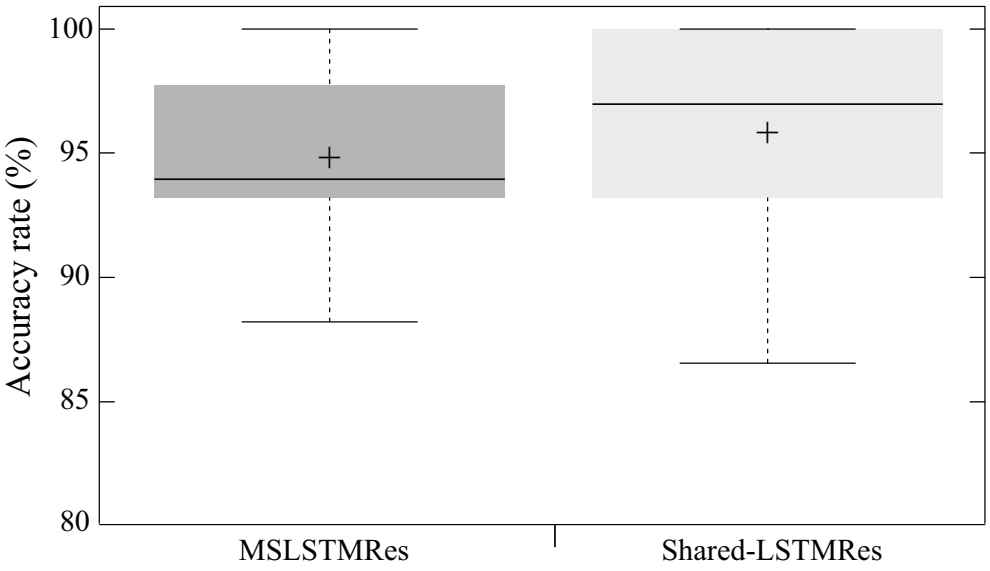


Fig. 3.6: Average accuracy rate of the proposed model. Comparison between MSLSTMRes (94.69 %) and MV-DNN (95.70%).

D Score Fusion

This experiment compared the proposed approach's performance when using features and score fusions. Features fusion estimates action probability using a combination of the features from cameras. Scores fusion, however, combined the prediction scores from multi-view inputs using the related arithmetic or geometric mean.

This experiment used pre-trained VGG-16 and shared-weight LSTMRes as CNN and RNN blocks for the proposed model, respectively. The models were trained with scenario II (Section Pre-processing and Learning)

The proposed DNN model exhibited an average accuracy of 97.22 % with the

arithmetic mean on IXMAS (Fig 3.7), which was 1.51% higher than the DNN model with feature fusion. Scores fusion with the geometric mean created the opposite effect, decreasing the proposed model's accuracy rate by 1.77 %. These results suggested the DNN model performed better when using the arithmetic mean as the score fusion.

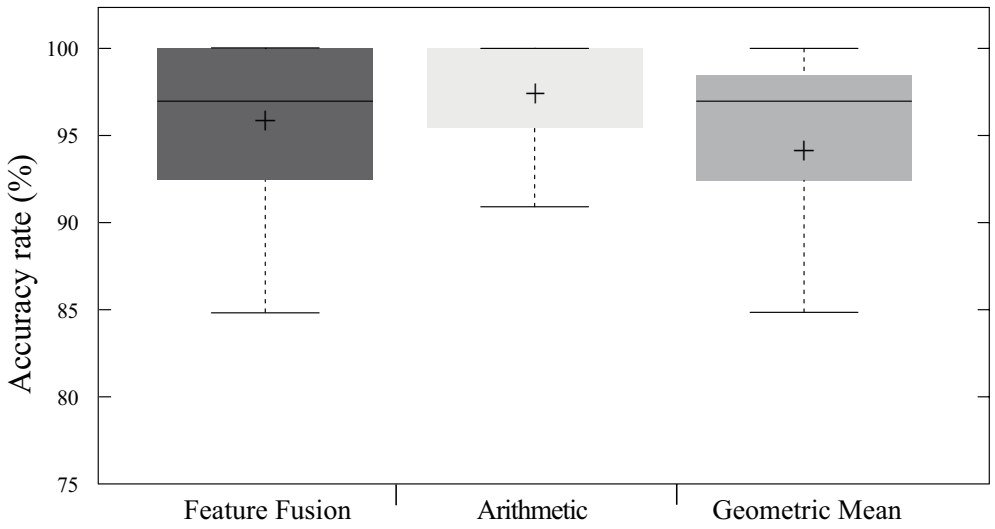


Fig. 3.7: Average accuracy rate of the proposed model using LSTMRes with feature-fusion and score-fusion techniques. Accuracy of shared-weights LSTMRes with feature fusion (95.71%), and score fusion employing the arithmetic mean (97.22%) and geometric mean (93.93%).

E Final Configuration

The above exploratory studies showed that the employment of VGG-16 (block5_pool), shared-weight LSTMRes, and shared fusion with the arithmetic mean significantly improved ($n = 396$, $p = 0.004$) the performance of the proposed model by $7.07 \pm 14.03\%$, compared to the model in our previous work [1]. The highest improvement (4.29%) was observed with pre-trained VGG-16 as a CNN block and fine-tuning of its last layer (Table 3.1), while the lowest improvement (1.01%) was gained when MSLSTMRes was replaced with shared LSTMRes. Although higher-resolution than in previous research [1] was used here (128 x 128 vs 73 x

73 pixels), the recognition rate with the proposed model increased by only 0.25%. Such improvement is insignificant, considering the small dataset.

Table 3.1: Average accuracy gain (%)with new configurations.

Configuration	Accuracy	Improvement
Previous work [1]	90.15	-
128 x 128 image	90.40	0.25
Pre-trained VGG-16	94.69	**4.29**
Shared-LSTMRes	95.70	1.01
Score fusion	97.22	1.52
Total		7.07

Application of the optimized configuration also increased the accuracy of the proposed model in recognizing actions performed by hands (e.g., watching check, arms crossing, waving, and punching) (Fig 3.8). The highest improvement (25%) was achieved in the identification of waving. We used the aforementioned new configuration for the next experiments: comparison of the proposed model with a single-input model and the state-of-the-art methods.

3.4.2 Single-view Classification

We performed an experiment on IXMAS with 13 actions to investigate the proposed model's performance using multi-view and single-view inputs. Comparison results (Fig 3.11) of the model using multi-view inputs and single-view inputs demonstrated that combining information from multi-view resulted in significant ($n = 396$, average $p < 0.05$) higher improvement. Also, the results signified that using inputs from **Cam5** caused the model to produce a 37.18% lower recognition rate while using information from the other views yielded a $15.92 \pm 1.23\%$ lower accuracy rate.

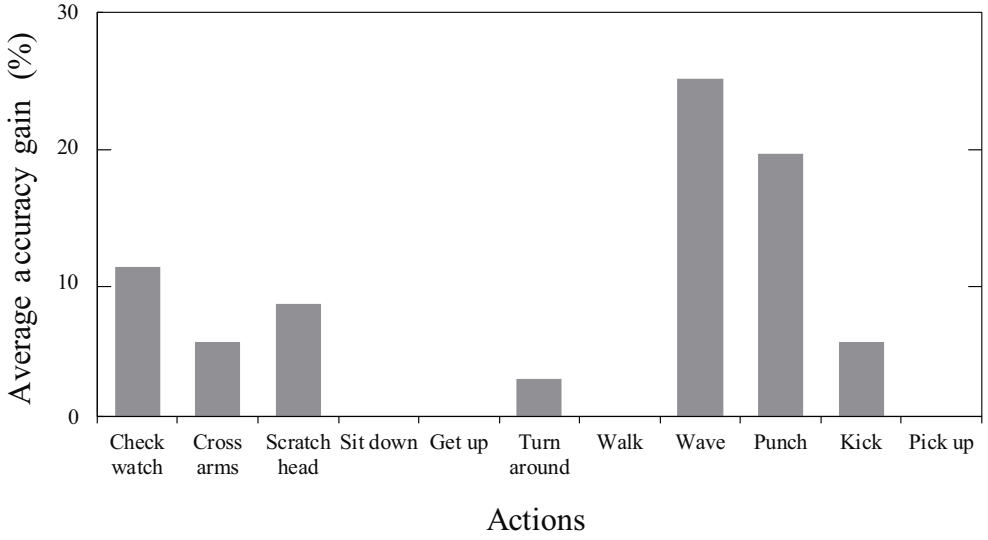


Fig. 3.8: Improvement of recognition rate with the revised model for each class. There was no improvement in recognizing the-"sit down/get up/pick up"-actions, as perfect recognition rate was achieved by the model with the previous structure. The highest accuracy gain was in recognizing wave action (25%).

3.4.3 Comparison with State-of-the-art Methods

We compared the proposed model to state-of-the-art methods on IXMAS and i3DPost (Table 3.2 and 3.3). Note that their results were not reproduced and the proposed model used 2D RGB images as inputs. Following the previous studies' experiment protocol, we evaluated the proposed model on the IXMAS dataset, with 11 subjects performing 10-action. We used data of all subjects in the evaluation of 13 actions on IXMAS and 10 and 12 actions on i3DPost. This experiment used learning scenario III (Sec. Pre-processing and Learning).

In evaluation with IXMAS,the proposed model significantly outperformed all 2D methods in recognizing 11 actions by 12.05% on average (Table 3.2). Performance was higher than with 3D methods [46, 62] but was slightly lower than the outcomes reported in [63]. Also, the proposed model got a 4.46% higher accuracy rate than other DNN models using 2D inputs and achieved competitive results to the models using 2D + optical flow inputs. Yet, the accuracy rate of proposed

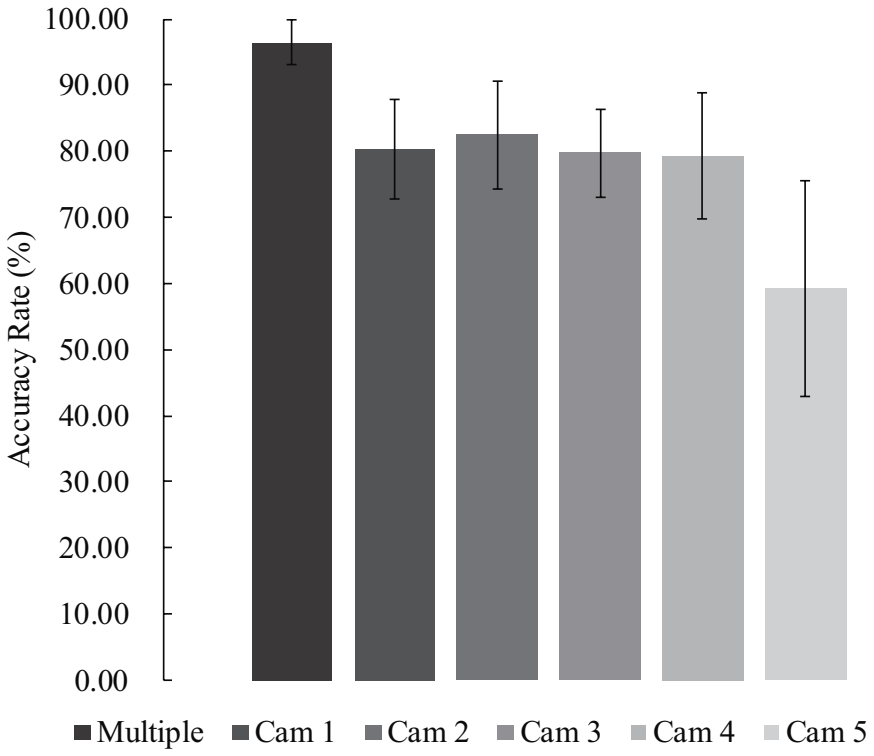


Fig. 3.9: Comparison between multi-view and single-view approaches. Recognition rate of proposed model utilizing multi-view inputs ($93.67 \pm 3.39\%$) and single-view inputs from Cam 1 ($80.34 \pm 7.57\%$), Cam 2 ($82.487 \pm 8.10\%$), Cam 3 ($79.70 \pm 6.66\%$), Cam 4 ($79.27 \pm 9.56\%$), and Cam 5 ($59.19 \pm 16.37\%$).

model was 2.33% lower than that of the methods employing adaptive score fusion.

In addition, the model produced a recognition of 96.37% in classifying 13 actions, outperforming Pehlivan *et al.* [62] with the use of 3D features. However, the proposed DNN model's recognition rate was still lower than 4D models [64].

The performance of the proposed DNN model in recognizing 10 actions on i3DPost was comparable to state-of-the-art methods (Table 3.3). The proposed model often misclassified actions with similar body configurations such as jumping and bending and exhibited confusion with differentiation of single and combined actions, such as “walking” and “running-jumping-walking” (Fig 3.10). The model achieved higher performance in classifying 10-action and 2- interaction.

Table 3.2: Comparison for recognition (%) using the proposed model with state-of-the-art methods on IXMAS .

Method	**Input**	**11 Actions**	**13 Actions**
Holte *et al.* [64]	4D	100.00	100.00
Turaga *et al.* [63]	3D	98.78	-
Spurlock *et al.* [65]	Dynamic	94.24	-
Weinland *et al.* [46]	3D	93.30	-
Pehlivan *et al.* [62]	3D	90.91	88.63
Vitaladevuni [66]	2D	87.00	-
Chaaraoui *et al.* [34]	2D	85.86	-
Liu *et al.* [67]	2D	82.80	-
Khan *et al.* [39] *	3D	99.60	-
Gao *et al.* [35] *	2D + optical flow	99.60	-
Purwanto *et al.* [41] *	2D + optical flow	97.22	-
Gnouma *et al.* [40] *	2D	92.81	-
Proposed model	2D	**97.27**	**96.37**

The "Input"-column indicates type of features used in the approaches. Khan *et al.* [39] utilized 50:50 training and test evaluation method, while Gnouma *et al.* [40] merely evaluated their model with 10 actions. * indicates DNN based approach or the methods employing CNN based features.

The proposed model got an average F1-score higher than 0.9 for all classes with all datasets (Table 3.4). The proposed model achieved the lowest F1-score when evaluated with 10-action of i3DPost while testing it with 11-action of IXMAS yielded the highest F1-score.

3.4.4 Online Classification

In the online scenario, we did not segment individual action sequences via the action labels as described in Sections Exploration Studies and Comparison

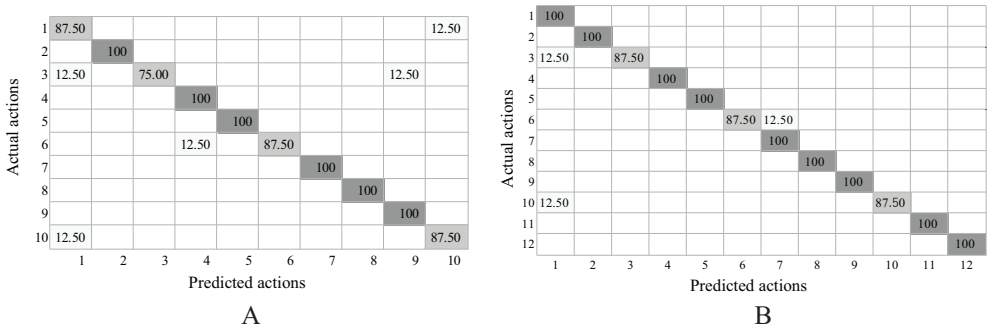


Fig. 3.10: Average accuracy rate of proposed model on i3DPost. The row and column represents action: walk(1), run(2), jump(3), bend(4), hand-wave(5), jump-in-place(6), sit-stand up(7), run-fall(8), walk-sit(9), run-jump-walk(10), hand-shake(11), pull(12). A and B illustrates experimental result on 10 and 12 actions, respectively.

Table 3.3: Comparison for recognition (%) of proposed model with state-of-the-art methods on i3DPost .

Method	**Input**	**10 Actions**	**12 Actions**
Spurlock *et al.* [65]	Dynamic	97.65	-
Holte *et al.* [64]	4D	97.50	-
Kose *et al.* [68]	3D	95.50	-
Tran *et al.* [69] *	3D	96.70	-
Mygdalis *et al.* [70] *	3D	95.51	-
Angelini *et al.* [71] *	Skeleton	99.47	-
Proposed model	2D	**93.75** (75/80)	**96.87** (93/96)

The "Input"-column shows the type of features used in the methods.
Evaluation was performed with the proposed model based on actions performed by one subject and two subjects ("12 Actions"-column). Mygdalis *et al.* [70] validated their model's performance using 3-fold cross-validation. * indicates DNN based approach or the methods employing DNN based features.

with State-of-the-art Methods. Rather, a sliding window was used to create N clips from video content. The proposed model should determine early and

Table 3.4: Average F1-score of the proposed model with 11 and 13 actions of IXMAS, and 10 and 12 actions of i3DPost.

	F1-Score (mean ± S.D.)
IXMAS-11	0.975 ± 0.026
IXMAS-13	0.963 ± 0.025
i3DPost-10	0.937 ± 0.062
i3DPost-12	0.969 ± 0.038

ambiguous actions (Fig 3.11) from unfinished sequences of actions or transitions phases between actions. To clarify how much information is required for the proposed model to make an accurate prediction, variable sliding time window t was used with the values of 10, 20, 30, 40, and 50. We trained the proposed model using learning scenario II (Section Pre-processing and Learning) for recognition actions with each time step. The final prediction was the average probability scores overtime.

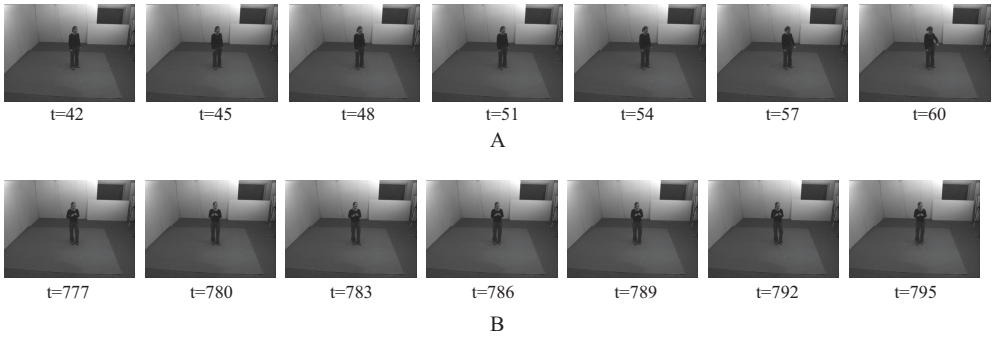


Fig. 3.11: Example of ambiguous-action clips. A: sequence of images from early watch-checking action. B: sequence of ambiguous actions (transition from punching to kicking action).

The experimental results (Fig 3.13) show that the highest accuracy and F1-score were got with t = 50. The accuracy and F1-score of the proposed model increased with longer sliding window values. However, this did not represent a proportional correlation, as the recognition rate was 0.69% lower at t = 20 than at t = 10.

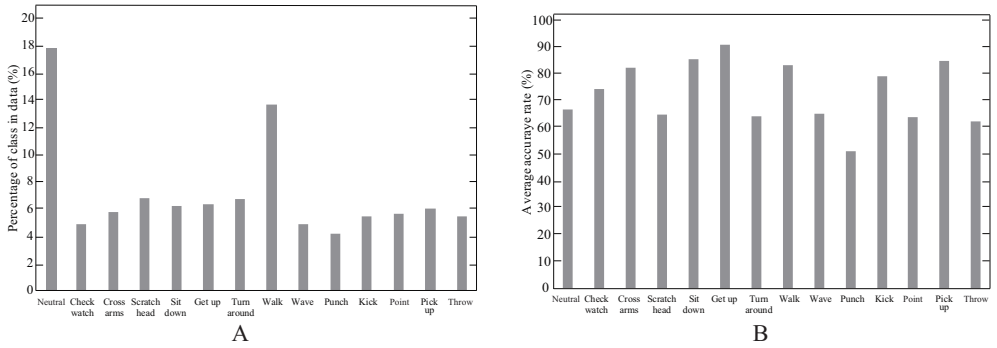


Fig. 3.12: Percentage labels in dataset and accuracy rate of the proposed model. A: percentage of classes in IXMAS dataset segmented with t equaled to 50. B: accuracy of the proposed model for each class with t = 50.

The imbalance dataset (Fig 3.12 (A)) did not impair the overall performance of the proposed model: the proposed model achieved F1 scores higher than 0.6 in all scenarios. Besides, the proposed model classified sitting-down, getting-up, and picking-up actions with over 80% accuracy rate, even though the percentage of data based on such actions was lower than the others. Yet, the experimental results for t = 50 (Fig 3.12 (B)) shows issues with the proposed model in differentiating actions performed only by hands (e.g., head-scratching, waving, punching, pointing, and throwing), with a recognition rate at less than 70%.

3.5 Concluding Remarks

This study presents a novel DNN method for estimating human activity through multiple-view inputs. Even though we trained the proposed model with only a few RGB-image datasets, experimental results showed it could achieve competitive results. Comparison to the-state-of-the-art showed that MV-DNN outperformed conventional CV and DNN-based method using 2D inputs and got comparable results to approaches using 3D and 4D inputs.

The results also confirmed the advantage of using multiple-view inputs to estimate human activity in dynamic environment, in which self or inter-object

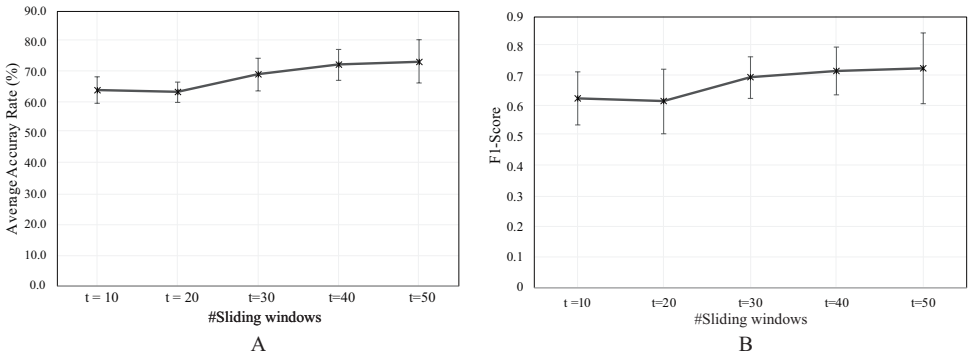


Fig. 3.13: Average accuracy rate of the proposed model in online classification. A: accuracy of the proposed model with a varying number of sliding windows: t = 10 (64.24 ± 4.26%); t = 20 (63.55 ± 3.45%); t = 30 (69.36 ± 5.43%), t = 40 (72.60 ± 5.15%), and t = 50 (73.64 ± 7.15%). B: average F1-score of the proposed model with a varying number of sliding windows: t = 10 (0.63 ± 0.08); t = 20 (0.62 ± 0.10); t = 30 (0.7 ± 0.06%), t = 40 (0.72 ± 0.07%), and t = 50 (0.73 ± 0.11%)

occlusion could occur. The proposed model attained higher accuracy rate when using multiple-view inputs than a single-view input. Similar to previous works' finding [31,32], this study found that combining information from multiple views resulted in a higher accuracy rate of the proposed model in MVHAR. That proved that additional information from another view could compromise the information loss caused by occlusion.

Despite its promising performance, the proposed model got an accuracy rate of less than 80% in online classification. Clips of transition from one action to another one caused ambiguous-action clips that impacted the proposed model's recognition rate.

Chapter 4

Modeling Behavior of ASD and Typical Children during Class Activity

4.1 Introduction

Previous works on ASD and ADHD children behavior have signified that disorder children behavior during play activities differed from their typical peers. Children with ADHD and ASD symptoms showed excessive physical movements, preference playing alone, and repetitive behavior [72].

Marker-less behavior monitoring system estimates children behavior without attaching markers or sensors to the children's body. Using tracking results, the system attempts to identify developmental disorder symptoms in children. Using a single camera, previous works have developed a system to monitor infants' general movement [16]. Yet, the method employing a single camera faced difficulty to monitor children behavior in nursery room during play activities because of occlusion [17].

This study employed multiple RGB cameras and Kinect sensors to track children behavior in the nursery room. Our proposed approach modeled the behavior

of children with PetriNet and represented each subject with four features. Statistical comparison was performed to find the difference between typical and ASD groups (Fig. 4.1).

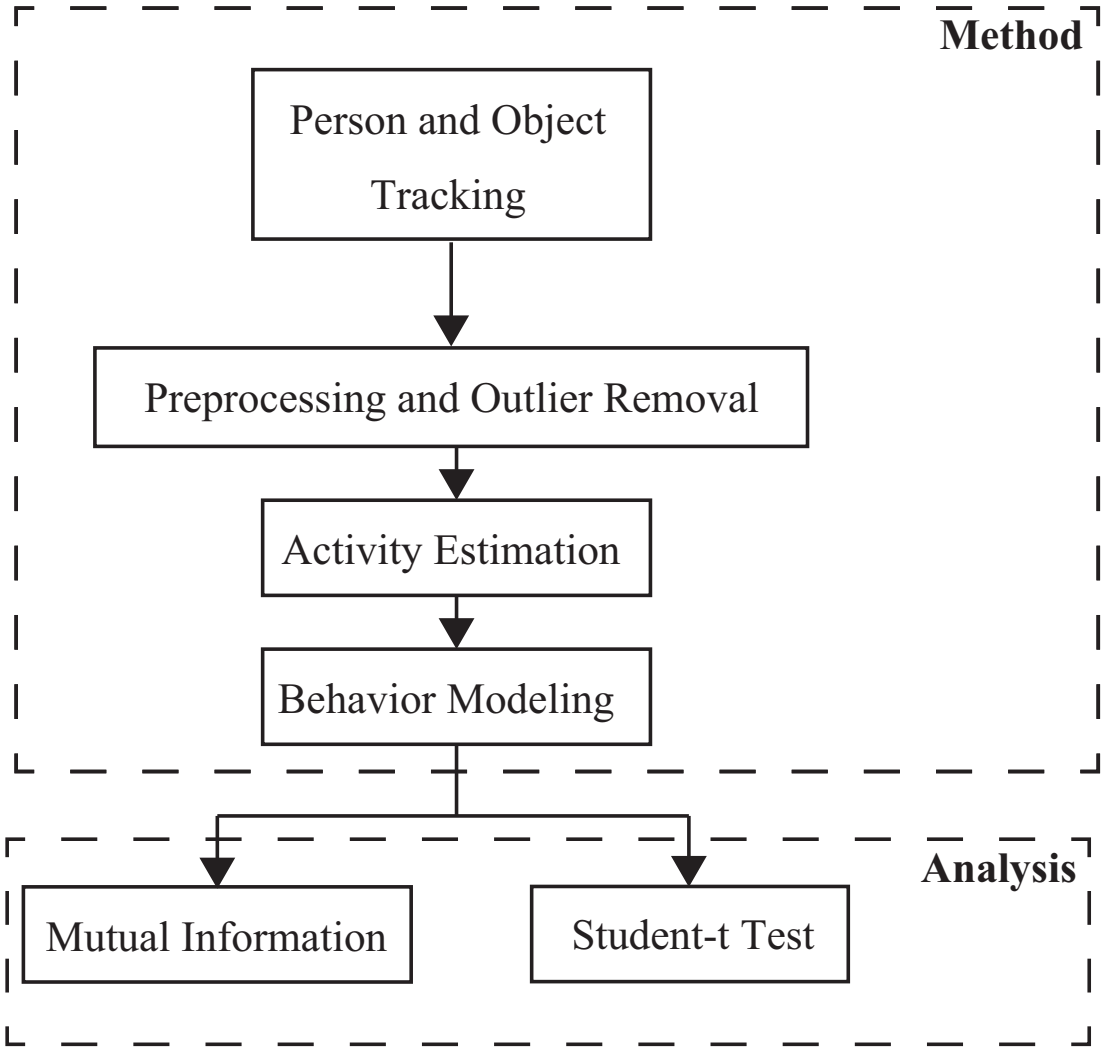


Fig. 4.1: Study flow diagram of Chapter 4.

4.2 Modeling Children's Behavior [2]

The proposed system comprised three steps: children and object tracking, children's activity estimation, and behavior modeling.

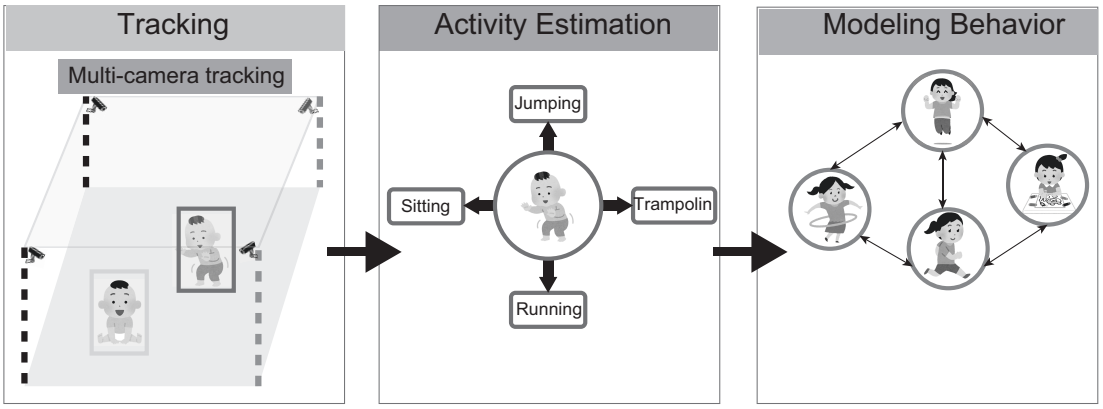


Fig. 4.2: Overview of behavior monitoring system proposed in this study.

4.2.1 Children and Object Tracking

The proposed model tracked positions of children and toys with OpenPTrack [73] (online-tracking) and OpenPose [74] (offline-tracking). OpenPose is an open source library that can track persons' skeleton in a real-time from a sequence of images. The API works by combining tracking results from multiple Kinect cameras to form 3D skeleton. In contrast, OpenPose detects human skeleton body from a single image and produces 2D joint-skeleton.

This study estimates objects' position with Yolo library [75]. Yolo divides an image into several grid cells, and then estimates class probability of each cell before combining them to segment objects. OpenPTrack combines Yolo's object detection results with depth information to locate the objects in 3D space.

4.2.2 Activity Estimation and Behavior Modeling

Our proposed system estimated children's activity using k-NN algorithm. First, the system computed centroid positions of children by averaging neck, left and right shoulder positions. Then, it calculated the Euclidean distance between a child's centroid and toys positions. The system assumed the closest toy to the child as the activity that was being performed by them.

PetriNet [76] (Fig. 4.3) was used to model the behavior of children. It con-

sisted of three components: playing states, transitions, and flows. For K activities and L external stimulus, the Petri net of a child's behavior was defined as $N = (P, I, I_0, T_p, T_i, T_0, F)$ where:

$$
\begin{aligned}
P &= \{P_1, P_2, P_3, \ldots, P_K\} \\
I &= \{I_1, I_2, I_3, \ldots, I_L\} \\
I_0 &= \{I_0\} \\
T_p &= \{T_{1,1}, T_{1,2}, \ldots, T_{k,k`}, \ldots, T_{K,K}\} \\
T_i &= \{T_{1,I_1}, \ldots, T_{K,I_L}, \ldots; T_{I_1,1}, T_{I_L,K}\} \\
T_0 &= \{T_{1,I_0}, \ldots, T_{K,I_0}; T_{I_0,1}, \ldots, T_{I_0,K}\} \\
F &\subseteq PT \cup IT \cup I_0T_0 \\
PT &= (P \times T_p) \cup (T_p \times P) \\
IT &= (I \times T_i) \cup (T_i \times I) \\
IT_0 &= (I_0 \times T_0) \cup (T_0 \times I_0)
\end{aligned}
\tag{4–1}
$$

Using action set $S_n(t) \in \{P, I\}$ and stimulus set $S_I(t) \in \{I, I_0\}$ from the network, the proposed system extracted five features from a child: their frequency to change (H_n) and to perform (R_k^n) activity; the average number of other children in the same state (F_k); their frequency to play alone (A_n); and their tendency to follow external stimulus (P_i). For the T time-step, those features were computed using the following formulas:

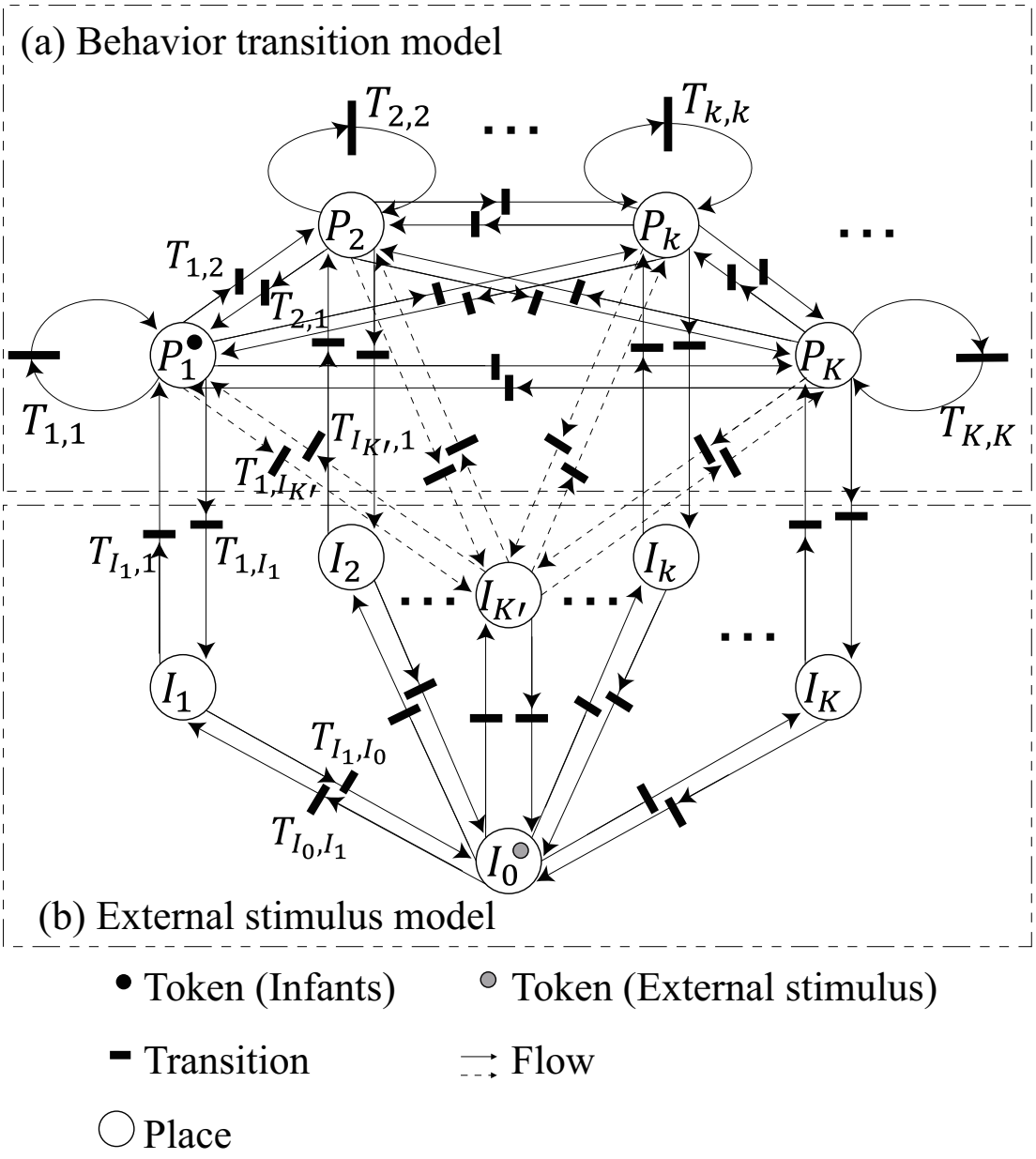


Fig. 4.3: PetriNet used in the proposed system to model a child's behavior.

$$
\begin{aligned}
H_n &= \int_0^T f(S_n(t), S_n(t+1))dt \\
R_k^n &= \int_0^T \bar{f}(S_n(t), P_k)dt \\
F_n &= \frac{1}{T}\int_0^T \sum_{i=0, i\neq n}^{N} \bar{f}(S_n(t), S_i(t))dt \\
P_I^n &= \frac{1}{T}\int_0^T \bar{f}(S_I(t), S_n(t))dt \\
A_n &= \frac{1}{T}\int_0^T a_t^n dt
\end{aligned}
\tag{4–2}
$$

where,

$$f(s,s') = \begin{cases} 1 & s \neq s' \\ 0 & s = s' \end{cases}$$
$$a_t^n = \begin{cases} 1 & \sum_{i=0,i\neq n}^{N} \bar{f}(S_n(t), S_i(t)) = 0 \\ 0 & others \end{cases} \tag{4–3}$$

and $\bar{f}(s,s')$ is complement of $f(s,s')$.

4.3 Method

Participants

Four male and two female disorder preschooler children participated in this study. The children were recruited from a special nursery school in Hiroshima prefecture, Japan. Exclusion criteria included any physical disorder and disabilities. All children had been diagnosed with having Autism Spectrum Disorder (ASD) and attention deficit symptoms. Before the experiment, subjects took Developmental Quotient (DQ) test and their score ranged from 85 to 112.

The control group included five males and four females typical preschooler children that did not have any physical and psychological disorder. We requited them from another nursery school in Hiroshima prefecture, Japan. The experiments for disorder and typical children were conducted separately in their nursery school.

4.3.1 Preprocessing and Data Analysis

A Mutual information

Mutual information (MI) [77] measures mutual dependence between two random variables (X and Y), in which the value is non-negative. High MI signifies that knowing the value of X determines the value of Y, while a zero value of MI means X and Y are independent.

Mutual information of two variables can be got by estimating the Kullback-Leibler distance D_{KL} between the joint probability $P(X, Y)$ and the product of marginal probability $P(X)P(Y)$, which is written as:

$$I(X;Y) = D_{KL}(P(X,Y)|P(X) \otimes P(Y)) \tag{4–4}$$

And for $X = \{x_1, x_2, \ldots, x_N\}$ the MI of each feature is computed as the following equation:

$$I(x_n;Y) = \frac{I(x_n;Y)}{\sum_{i=1}^{N} I(x_i,Y)} \tag{4–5}$$

The value of mutual dependence ranges from 0 to 1, in which 1 shows strong dependence between X and Y, while 0 implies that they are independent. This paper presents the mutual dependence score on a percentage scale.

4.3.2 Experiment Protocol

In this study, we conducted two experiments (Fig. 4.4) involving nine typical and six disorder children. Typical and disorder groups took part in the former and later experiments, respectively. The first experiment used four network cameras to record children's activity in a nursery room. The activity states comprised puzzle, hula-hoop, mat, and trampoline. We performed person and object tracking in an offline manner using OpenPose [74] and Yolo [75] for every single camera.

Different to the first experiment, we used not only network cameras but also OpenPtrack with three Kinect V2 to track children's pose and object's position in an online manner (Fig. 4.5). Tracking results from OpenPTrack [73] were streamed from UDP port and saved locally in the master PC. The activity states in this experiment comprised trampoline, slide, and puzzles.

Although the activity of a child was estimated for each frame, we assigned the activity label at time k by considering the maximum occurrence of activity within 30 seconds using the following formula:

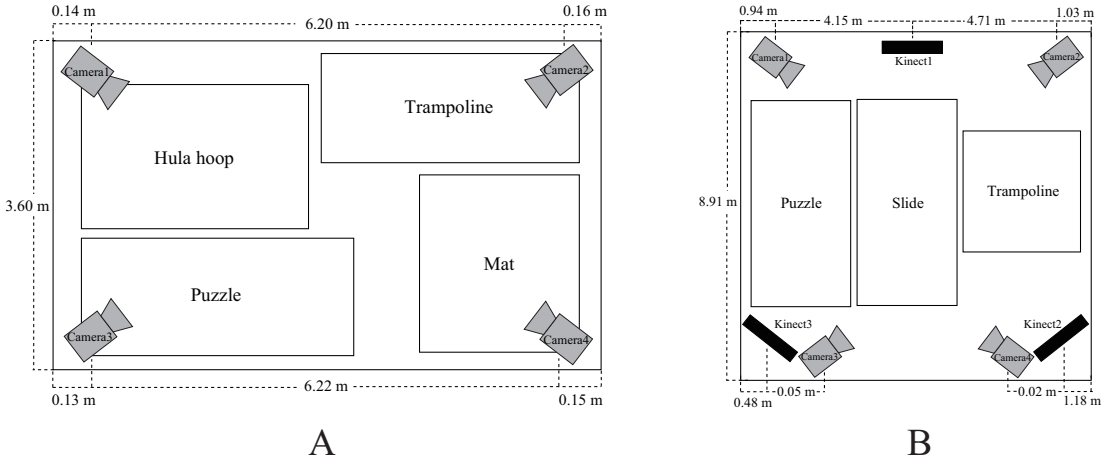


Fig. 4.4: The study conducted the first (A) and second (B) experiments in two different schools with different environments.

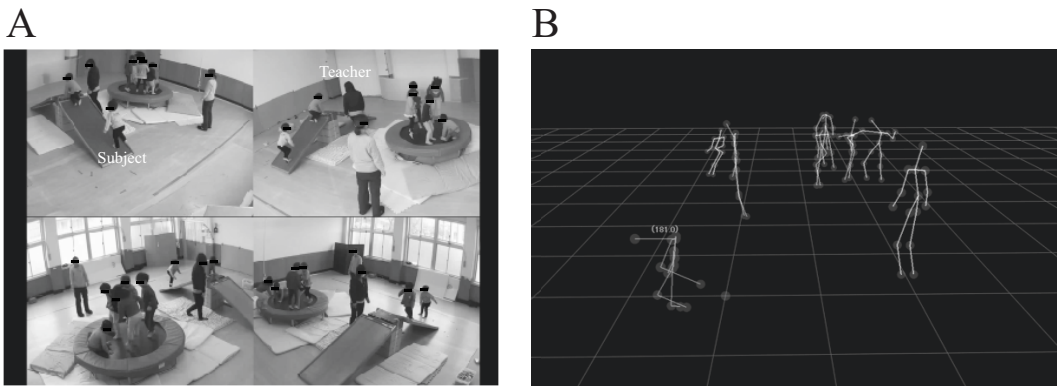


Fig. 4.5: (A) Recording results of multiple RGB cameras. (B) Tracking results of OpenPTrack with multiple Kinect sensors.

$$P_k = \underset{x}{\operatorname{argmax}}\{y_n \in Y | x \in y_n\} \tag{4–6}$$

where $y_n = \{y_{k+1}, y_{k+2}, \ldots, y_{k+30}\}$.

In both experiments, we categorized the activity into two types: dynamic and static. The static activity of the first and the second experiments was the puzzle, and the others were classified as dynamic. The ratio of a child to perform static activity over dynamic activity (S_n) was defined as:

$$S_n = \frac{R^n_{static}}{\frac{1}{N}\sum_{i=1}^{N} R^n_{dynamic_i}} \tag{4–7}$$

4.4 Results

4.4.1 Statistical Analysis

Mutual information scores suggested that the frequency of changing activity and the duration of playing alone were more informative than the average number of children in the same state and the frequency of changing activity to differentiate typical from ASD children.

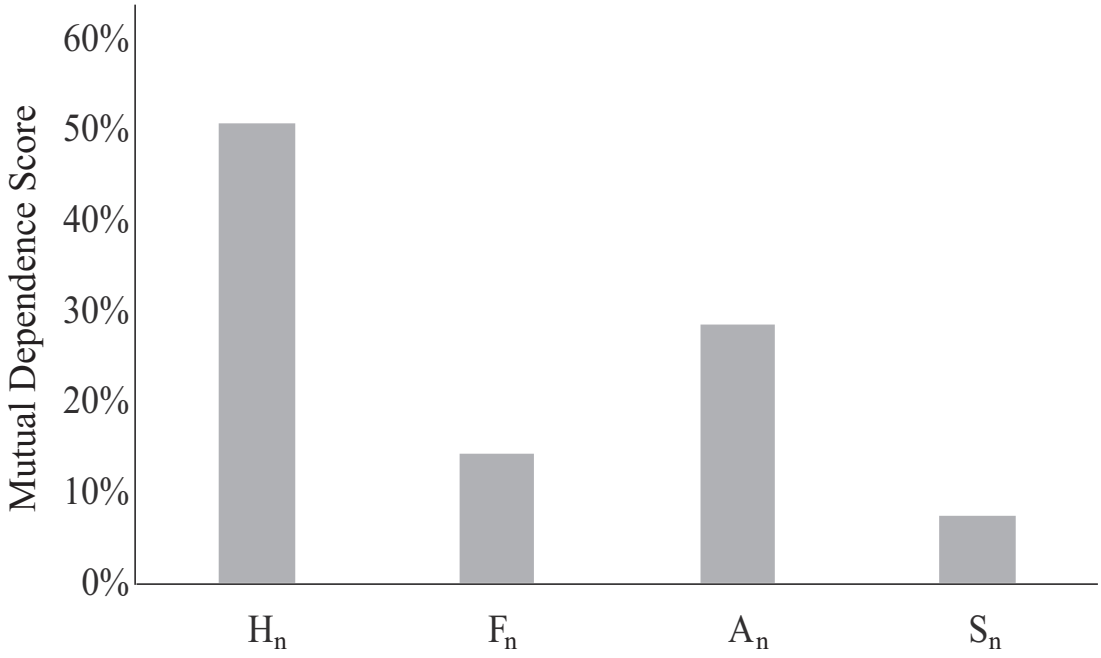


Fig. 4.6: Mutual dependence between children's behavior features and children's diagnosis results. Behavior features comprises the frequency of changing activity (H_n: 50.72%), the average number of children in the same state (F_n: 14.24%), the duration of playing alone (A_n: 28.51%), and the frequency of performing static activity (S_n: 7.37%).

The statistical comparison results agreed with the mutual information scores.

The results suggested a significant difference ($p < 0.05$) between typical and ASD groups for H_n and A_n by both Student t and MannWhitney U tests. Also, MannWhitney U test signified that ASD children showed solitude behavior; the p-value of $H_n < 0.05$.

Table 4.1: Statistical comparison with Student t and Mann-Whitney U tests. Behavior features comprise the frequency of changing activity (H_n), the average number of children in the same state (F_n), the duration of playing alone (A_n), and the frequency of performing static activity (S_n).

#	Mean		S.D.		Student	MannWhitney
	Typical	ASD	Typical	ASD	p	p
H_n	-0.021	3.567	0.451	1.473	0.000	0.001
F_n	-0.145	-0.291	0.178	0.066	0.080	0.034
A_n	-0.104	2.508	0.644	2.567	0.011	0.002
S_n	-0.404	0.284	0.155	1.174	0.100	0.069

The corresponding mean values of those significantly different variables suggested that children with ASD symptoms showed a tendency to change their activity, play alone, and separate themselves from others more often than the typical group.

4.5 Concluding Remarks

The study investigated whether it is possible to identify ASD disorder symptoms in children by employing tracking results from multiple camera. The proposed system tracked children's activity in the nursery room with multiple RGB cameras and Kinect sensors. Using PetriNet, the behavior of typical and ASD children was modeled and four features were extracted from it.

Statistical comparison between the groups and mutual information scores signified that ASD children had higher tendency to play alone and change their activity. These results agree with previous findings [78] that stated children with

ASD disorder symptoms change their activity more frequently than the control group and have little interest in peers.

Two limitations in this study are the accuracy of tracking results and the separate environments used in the experiments. We realized that involving over seven subjects in the experiment impaired OpenPTrack's performance employing three Kinects that forced us to perform tracking manually. Second, we did not change the playing room during experiments to get the natural behavior of children. While it allowed us to minimize the possibility of the children altering their behavior intentionally, it also introduced bias in our results since the groups did not perform activities in the same environments. Future studies may experiment with different protocols: instead of measuring children's natural behavior, they can evaluate how children react and adapt to a new environment. Also, future experiment should consider the number of cameras must be proportionate to the number of children in one session.

Chapter 5

Investigation of Response and Gaze Behavior during the Go/NoGo Task

5.1 Introduction

Studies on performance during the Go/NoGo task of children with ADHD and ASD symptoms have discovered a significant difference between those children's response and their typical peers [4, 10, 79]. Higher response time variance was found to be related to both of disorders and might be caused by variability in neural activation of children with the disorders [80].

Previous studies of ASD children have observed that gaze-adjustment of children with ASD symptoms was slower during eye-tracking measurement of joint attention [13], and more irregular during face-to-face conversation [14]. They also have found that temporal features from gaze modulation were more informative than global-spatial features in identifying ASD symptoms on children.

Using features from gaze movement, other studies extend those works by utilizing machine learning to differentiate automatically typical from ASD children. They asked participants to take part in face-to-face conversation [81] or to com-

plete visual tasks such as viewing a sequence of face images [82] or identifying directional cues [83]. Then they used spatial features extracted from children's eye movement distribution to recognize ASD symptoms in children.

This study aims to measure children's response and gaze behavior during Go/NoGo task. We developed a game version of Go/NoGo task (CatChicken game) that measure children's response and gaze behavior through spacebar and an eye-tracker, respectively. Afterwards, we extracted features from their motor response and gaze behavior and performed statistical analysis to identify informative features (Fig. 5.1).

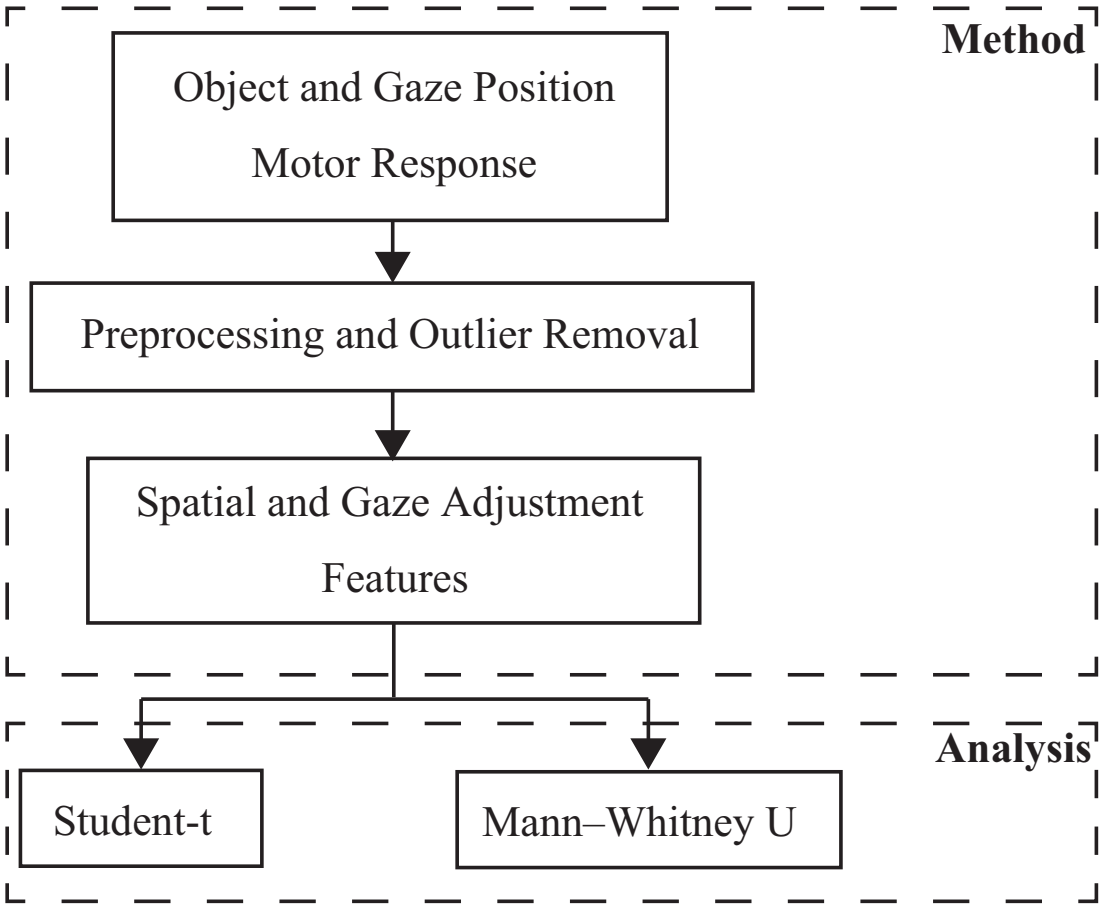


Fig. 5.1: Study flow diagram of Chapter 5.

5.2 A Serious Game Based on Go/NoGo Task [3]

The CatChicken system comprises training and evaluation functions (Fig. 5.2). Using the training function, an instructor can change the game's parameters that include the task duration, the appearance time of a stimulus, the proposition of stimuli, the interval between two consecutive stimuli, and the locations at which a stimulus appears. While using the evaluation function, those parameters can be fixed to create a standardized task for all subjects.

The game interface represents Go and NoGo stimuli as "cat" and "chicken" characters, respectively. The game requires the subject to respond to a "chicken" character by pressing a spacebar and inhibit their action towards a "cat" character. A character can appear in one of nine locations represented by red flowers. The system outputted the user's response types and time, and stimulus and eye locations on the monitor (Fig. 5.3).

A user responded to the stimulus by pressing the spacebar. The system categorized a subject's response as one of four types: Go-positive if the subject responded to the Go character; Go-negative if he missed it; NoGo-positive if he inhibited his action in response to the NoGo character; NoGo-negative if he reacted to it. Different audio feedback was given when the subject responds correctly and incorrectly towards the stimulus. The system was equipped with a Tobii 4C eye tracker that recorded the user's eye position on the monitor continuously. The eye tracker sampling rate was 90Hz (interlaced), and its operating distance was 50 cm to 95 cm. Stimuli and eye locations on the monitor were normalized to the unit interval $[0, 1]$ by dividing the pixel coordinates by the window's coordinate length.

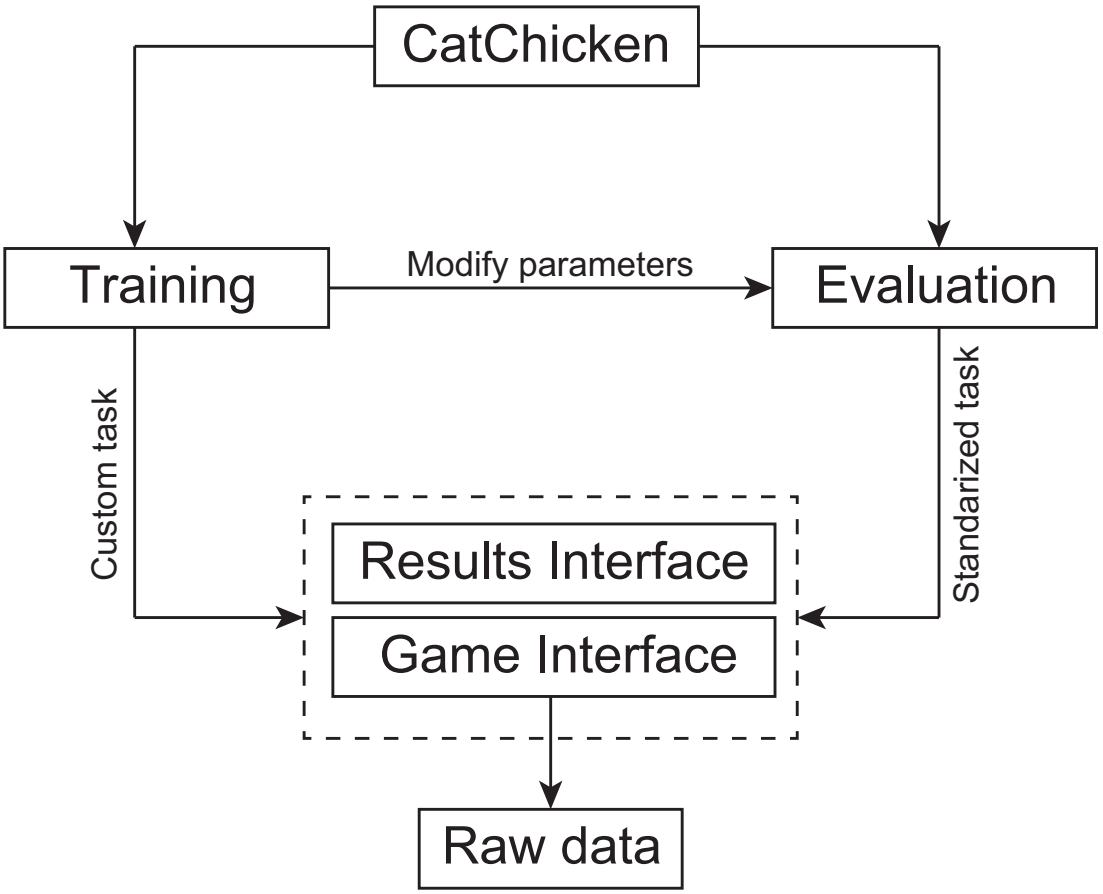


Fig. 5.2: Architecture of the CatChicken system. The game produces raw data that consist of the subject's response types and times, and locations of stimulus and gaze over time.

5.3 Relation Between Response and Gaze Behavior

This sections explains the results of our study in relation between subject's response and their gaze behavior. We developed a serious game version of Go/NoGo task and measured participants' response and gaze behavior using the system. Statistical and clustering analysis were conducted to understand the data patterns.

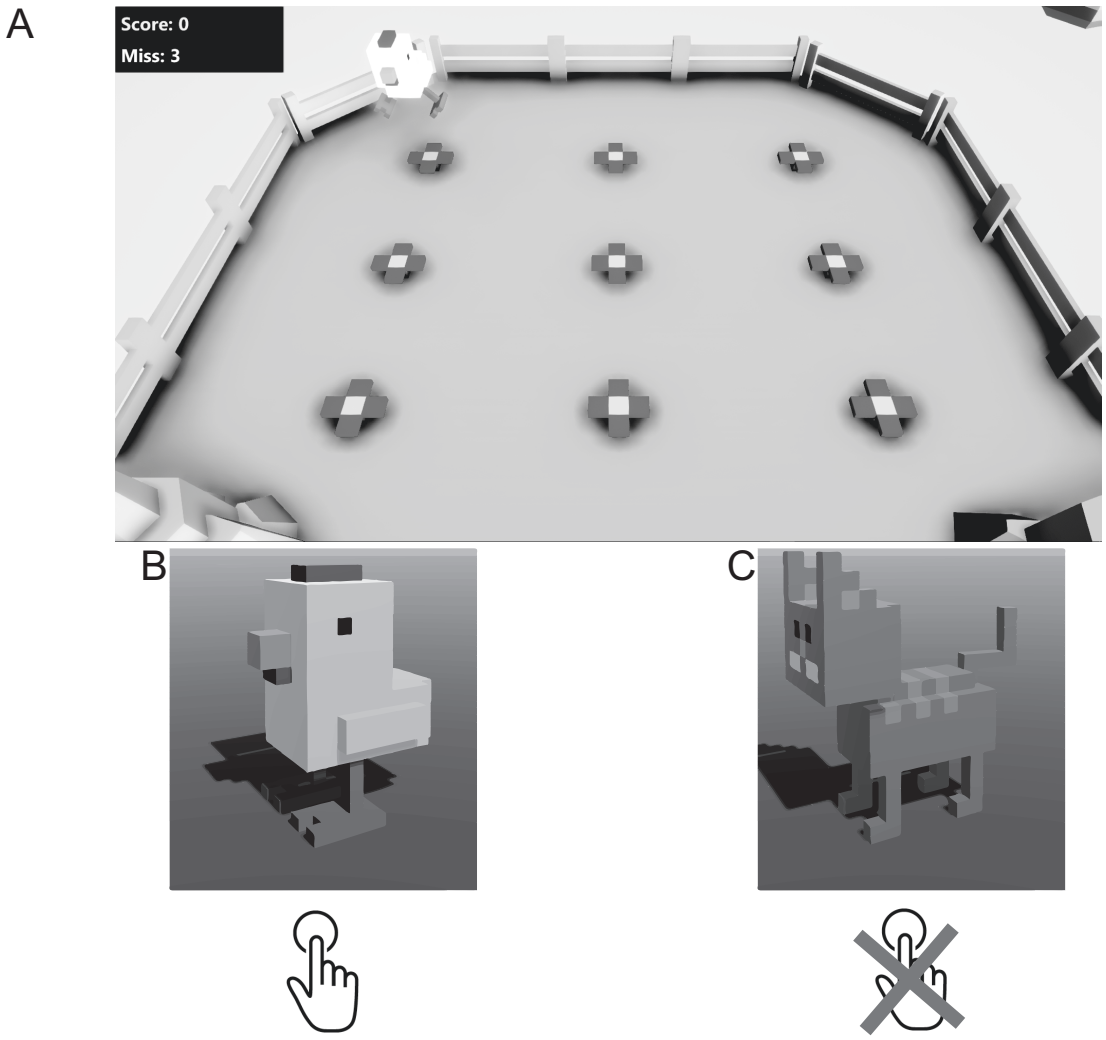


Fig. 5.3: Game interface of the CatChicken system. (**A**) Nine red flowers representing the locations in which a stimulus can appear; (**B**) Go and (**C**) NoGo characters.

5.3.1 Game Performance and Gaze Behavior Features

This study represents each participant with four response and two gaze-behavior features. Response features includes Go-error, NoGo-error, response time (RT), and variance of RT (RT-var). Go and NoGo error score were the frequency of a child to respond incorrectly towards Go and NoGo stimulus, respectively. RT was the average value of the difference between spawn time of

stimulus (t_s) and the participant's reaction time (t_{rc}). RT-var was the standard deviation of response times.

$$t_{rs} = t_a - t_r \tag{5–1}$$

Gaze behavior features comprised gaze trajectory area and gaze-to-object position. Gaze trajectory are was estimated with the Convex Hull algorithm; its value ranged from 0 to 1. The gaze-to-object position was the Euclidian distance between subjects' gaze position and the stimulus position when they responded to the stimulus.

5.3.2 Method

A Participants

We explained about the aim and procedure of this study to participants and asked their consents before starting the experiment. The study was approved by the Research Ethics Committee of the Prefectural University of Hiroshima (letter no: 15MH070) and was conducted under the amended Declaration of Helsinki.

We conducted two experiments in this study. The first involved university students and aimed to investigate the relation between subjects' response and gaze behavior. The second had an aim to identify the difference of response and gaze behavior between typical and ASD children.

In the first experiment, we recruited 59 university students from Hiroshima Prefecture and Yokohama National universities. The participants comprised 31 males and 28 females, with an average age of 25 $\pm$ 3.30 years. All participants did not have physical and mental disorders.

The experiment comprised four steps. First, an instructor calibrated eye-tracker for the subject and explained the game UI. Next, the subject fill in their personal information. Then, they took one-minute training under the instructor's supervision before participating in 10-minute evaluation. Finally, the instructor explained the detail of the experimental results.

The second experiment involved two typical (11 and 7 years) and one ASD child (3 years) to identify the difference of their response and gaze behavior. All subjects were male and did not have any physical disorders. This experiment followed the same procedure as the first experiment. However, all children took part in a 3-minute test instead of a 10-minute evaluation.

B Pre-processing and Statistical Analysis

Before performing statistical analysis, features were normalized with z-normalization. Linear correlation between game performance and gaze behavior features was calculated using Pearson correlation. The correlation coefficient and significant values were respectively represented as p and r.

This study also performed clustering with K-Means to identify patterns in participant's response time and gaze movement. The number of cluster(k) was set to {2, 3, 4, 5}; the clustering results were eveluated with Silhouette score. In the clustering process, each subject was represented by RT, RT-var, and gaze trajectory area.

5.3.3 Results

A Relation between Participants' Response and Gaze Behavior

Analysis results showed that statistically significant ($p < 0.05$) relationship existed between gaze trajectory area and Go-error percentage, RT, and RT-var (Table 5.1). All relationships were positive that mean higher gaze trajectory area yielded higher Go-error, RT, and RT-var.

The most significant relationship ($p = 0.002$) was between gaze trajectory area and Go-error percentage. Its correlation of determination (r^2) was 0.157, which meant that 15.7% of variation in gaze trajectory area could be explained by Go-error percentage. The relation was greater in male than female subjects. Also, the standard deviation showed that high variance of Go-error was presented in all subjects.

Table 5.1: Statistical analysis results of game performance and gaze behavior results. r and p stand for correlation coefficient and p-value. The variance (STD) was computed using standard deviation.

			Mean (± S.D.)	Gaze Area (r)	Gaze Area (p)
Error (%)					
	Go		0.35 ± 5.92	0.397	0.002
		M	0.23 ± 4.84	0.480	0.006
		F	0.48 ± 6.92	0.385	0.043
	NoGo		0.72 ± 8.44	0.201	0.126
		M	0.62 ± 7.85	0.301	0.099
		F	0.82 ± 9.04	0.124	0.528
RT					
	All		460 ± 28.6	0.304	0.019
		M	456 ± 27.3	0.308	0.092
		F	463 ± 29.6	0.290	0.135
	Go		460 ± 62.1		
	NoGo		386 ± 130.8		
RT Var					
	All		56.29 ± 13.38	0.354	0.006
		M	52.52 ± 12.31	0.467	0.008
		F	60.46 ± 13.28	0.223	0.254
Trajectory Area					
		M	0.47 ± 0.15		
		F	0.49 ± 0.14		

The correlation coefficients suggested that subjects with greater gaze area responded faster with higher variance. The results, however, suggested that significant relationship was more pronounced in male than female groups ($p > 0.05$).

Table 5.2: Silhouette score of clustering with different values of k

K	2	3	4	5
Silhouette score	0.33	0.27	0.26	0.26

Table 5.3: Statistical comparison between the first and second clusters for gaze trajectory area, response time(RT), and response time variance (RT-var).

	Gaze Area	RT	RT-var
Student-t	0.001	0.000	0.000

B Clustering Results

The silhouette scores of clustering demonstrated that the higher the value of k was, the lower the silhouette score became.The most optimal performance was achieved with k equaled to 2.

The clustering results showed 28 and 31 people belonged to the first and second clusters, respectively. A significant difference between those clusters was found in the values of gaze-trajectory area, RT, and RT-var features.

Members of the first cluster modulated their gaze on the middle of the screen (average of gaze area = 0.42), while those of the second cluster adjusted their gaze to stimulus position (average of gaze area = 0.53). The results also showed that people belong to the second cluster responded slower with higher variance.

5.3.4 Preliminary Results of an ASD Child

Line chart of RT and RT-var suggested that ASD child responded slower with higher variance than typical group; on average, ASD child responded 59 ms slower with 80 ms higher variance.

Gaze modulation of the ASD child was more dispersed than his typical peers (Fig. 5.7). Typical children's gaze trajectory area was 0.1 smaller than the ASD child's. Although the difference was insignificant, the plot showed that gaze modulation of typical children was more structured than that of ASD child.

The ASD child' RT and RT-var became more similar to the typical children's

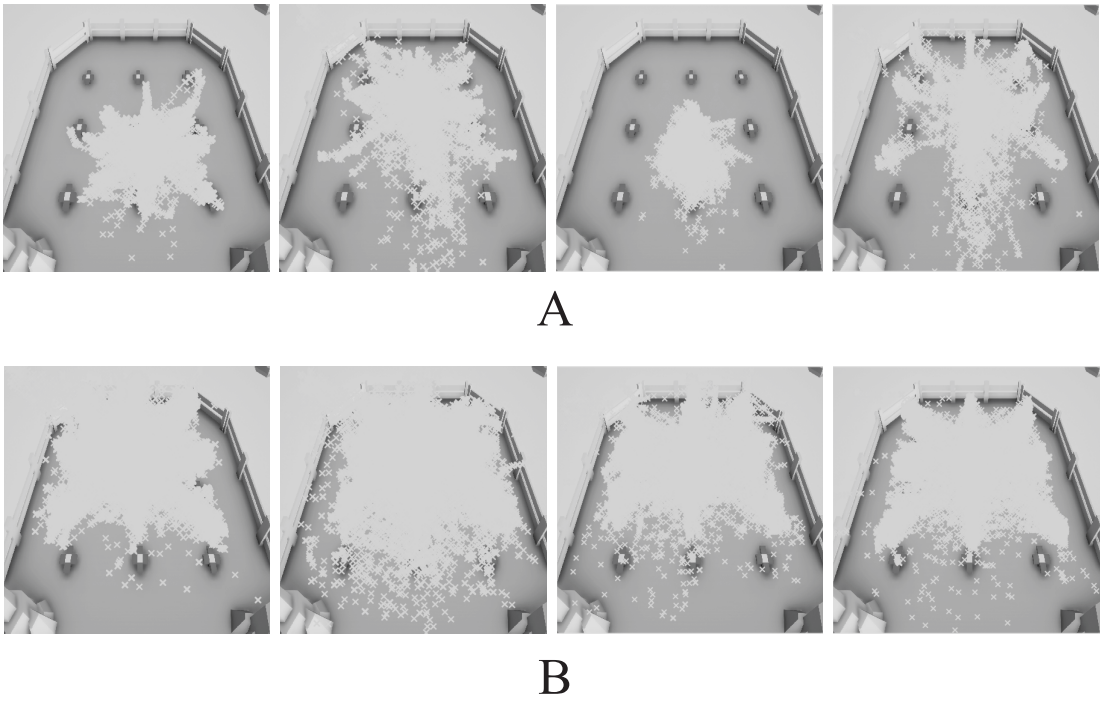


Fig. 5.4: Scatter plots of participants' gaze belonging to the first (A) and second (B) clusters. The gaze trajectory areas from left to right: first cluster are 0.28, 0.56, 0.17, and 0.65; second cluster are 0.61, 0.83, 0.73, and 0.67.

after he took part in rehabilitation. The child's response became ±100 ms faster after taking rehabilitation, yet more stable: the response time variability decreased ±146 ms after the second and the third treatment.

Similarly to RT and RT-var results, the ASD child's gaze modulation seemed to improve after the rehabilitation. His gaze trajectory area decreased after each rehabilitation; the average decrements was 0.23. The gaze pattern resembled that of typical children, which adjusted their gaze to the stimulus' position.

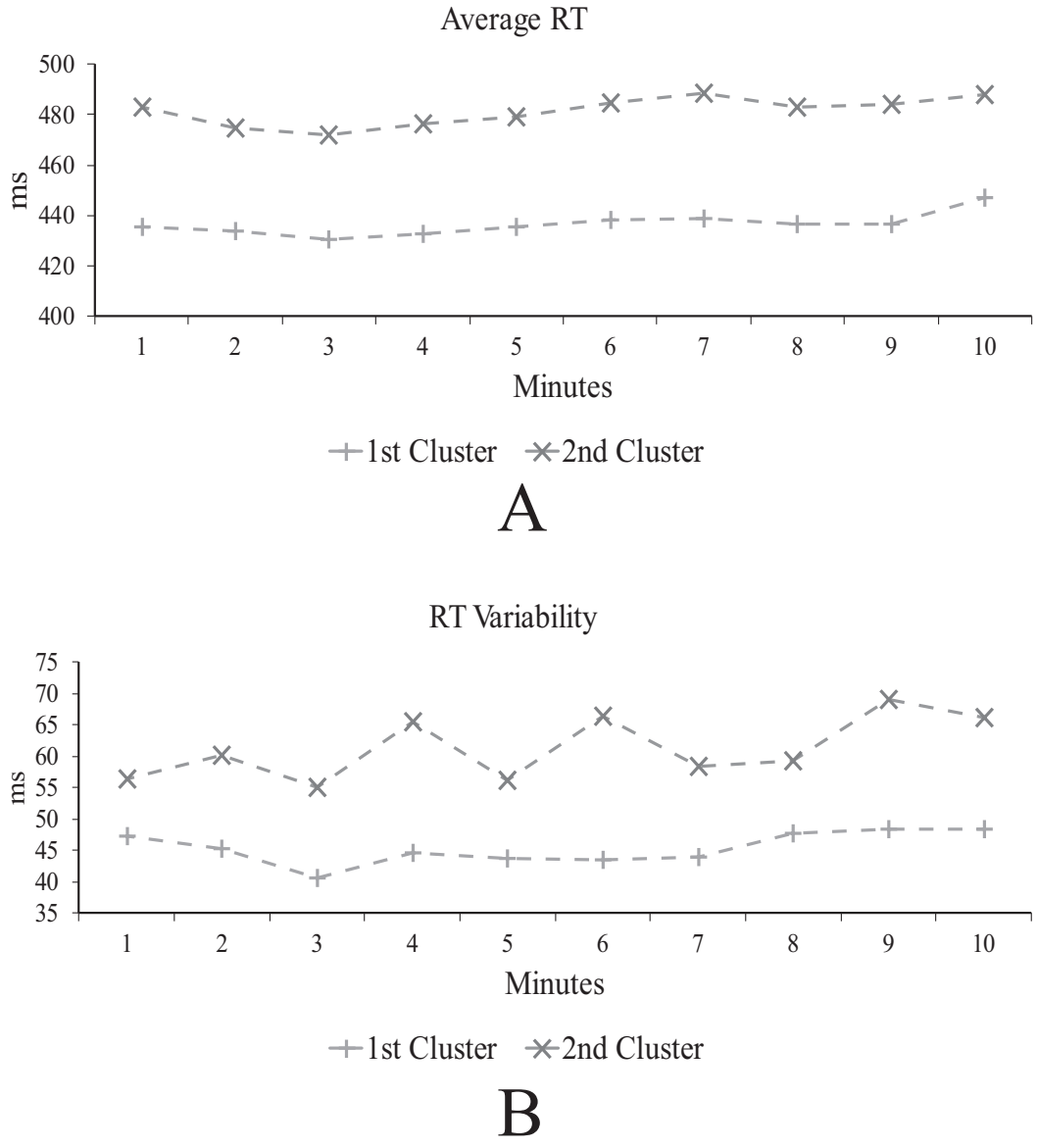


Fig. 5.5: Average (A) and variability (B) of the child's response time before and during rehabilitation.

5.4 Investigation of Response and Gaze Behavior of Children with ASD Symptoms [4]

This study aims to investigate the difference between response and gaze behavior of typical and ASD children during Go/NoGo task. Also, we conduct

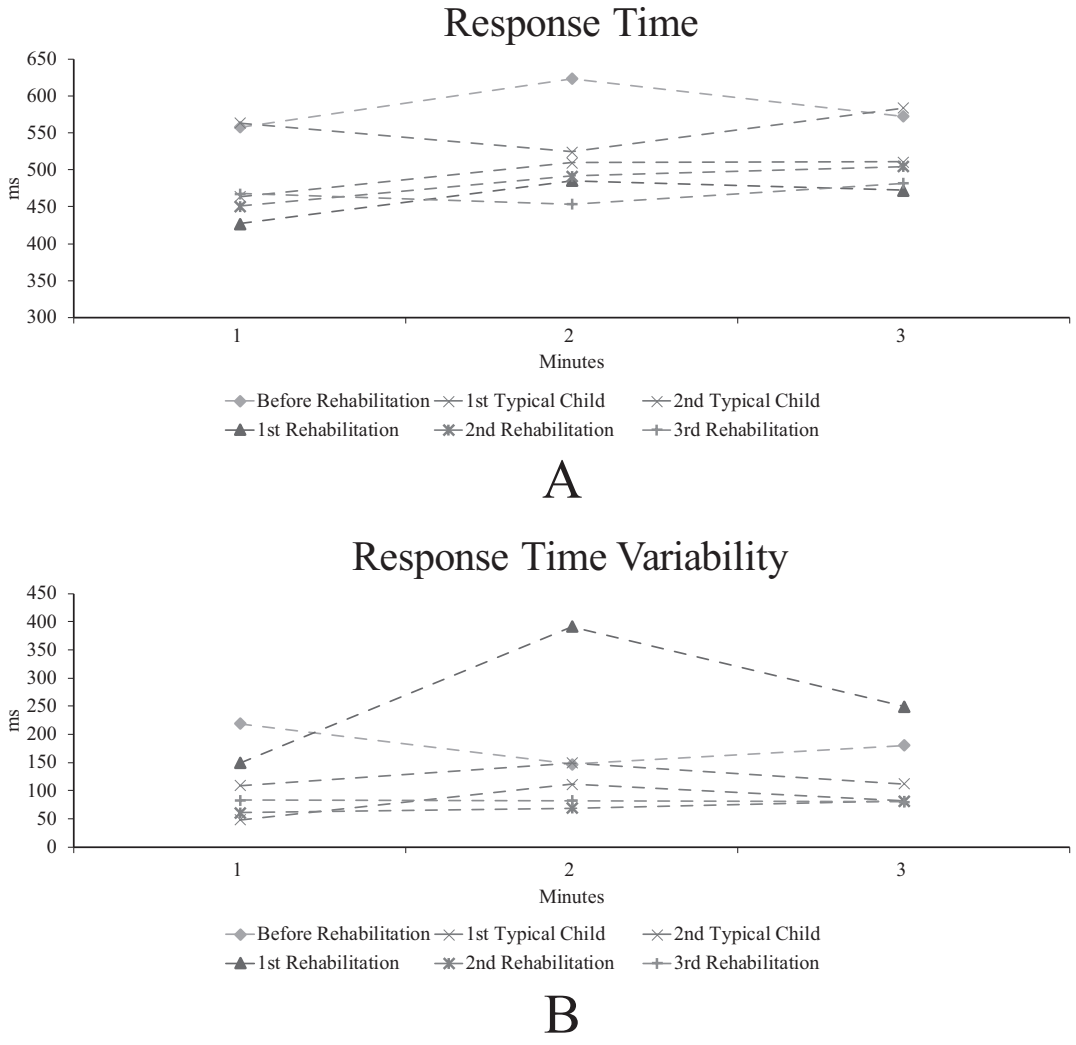


Fig. 5.6: Average (A) and variability (B) of subjects belong to first and second clusters

a statistical comparison of spatial and temporal features to find out the most informative features.

5.4.1 Features

Using the CatChicken game, we measured children's response and game performance during the Go/NoGo task. The game represented "Go" and "NoGo" stimulus as "Cat" and "Chicken" characters, respectively. The subjects re-

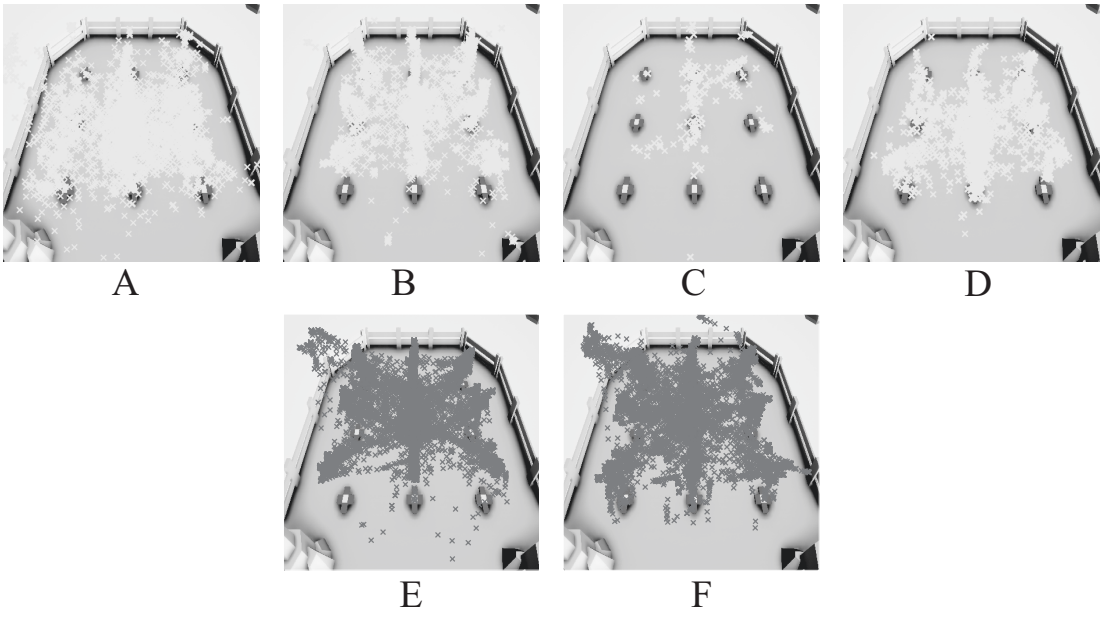


Fig. 5.7: An ASD child gaze trajectory: before rehabilitation (A), after first (B), second (C), and third (D) treatments. Gaze behavior of typical children: 11 year-old(E) and 7 year-old (F). The areas from left to right are 0.68, 0.64, 0.30, 0.36, 0.574 and 0.573.

sponded to the stimulus by pressing the space bar on the keyboard and the system tracked the participant's gaze movement with eye-tracker attached to the monitor.

The system outputted the user's response types and time, and stimulus and eye locations on the monitor (Fig. 5.8). Then using those information, the system extract spatial and gaze-adjustment features [4]. Spatial features included game performance, absolute gaze position, and gaze-to-object movement. Gaze-adjustment features measured the distance between participants' gaze and stimulus positions when the stimulus was presented onscreen.

A Spatial Features [4]

The spatial features consisted of 6 game performance, 10 absolute gaze position, and 8 gaze-to-object movement features (Tab 5.4. The features were com-

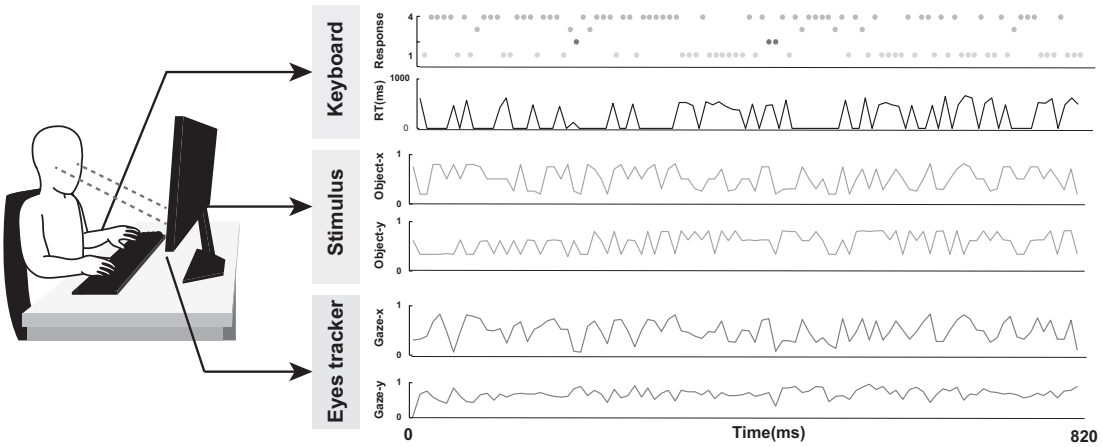


Fig. 5.8: Information measured by the CatChicken system. While playing the Go/NoGo game, CatChicken records children's response types and times, and locations of stimulus and gaze over time. The response types are Go-positive (green), NoGo-positive (blue), Go-negative (orange), and NoGo-negative (red). The values of object and gaze locations are normalized to range from 0 to 1.

puted using children's response types and times, and locations of stimulus and gaze.

Table 5.4: The list of spatial features extracted from response and gaze behavior: game performance, absolute gaze, and gaze-to-object movement features.

No	#	Detail
1	Go-positive	The percentage of Go response
2	Go-negative	The percentage of Go-negative response
3	NoGo-positive	The percentage of NoGo response
4	NoGo-negative	The percentage of NoGo-negative response
5	RT	The average of a subject response time
6	RT-var	The standard deviation of a subject response time
7	Trajectory-area	The gaze trajectory area
8	Velocity-avg	The average instantaneous velocity of subjects' gaze
9	Velocity-var	The standard deviation of the instantaneous velocity of subjects' gaze
10	Acceleration-avg	The average acceleration of subjects' gaze
11	Acceleration-var	The standard deviation of the velocity of subjects' gaze along the y-axis
12	Fixation-avg	The average of subjects' fixation time
13	Fixation-var	The standard deviation of subjects' fixation time
14	Distance-avg	The average of gaze distance
15	Distance-var	The standard deviation of gaze distance
16	Angle-avg	The average of gaze angle
17	Angle-var	The standard deviation of gaze angle
18	Distance-sen	Sample entropy of subjects' gaze distance
19	Angle-sen	Sample entropy of subjects' gaze angle
20	Velocity-sen	Sampe entropy of gaze velocity
21	Spatial-en	The entropy of subjects' gaze
22	Gaze-obj-en	The entropy of the distance between subjects' gaze and stimulus position
23	Gaze-obj-sen	Sample entropy of the distance between subjects' gaze and stimulus position
24	Gaze-obj-spe	Spectral entropy of the distance between subjects' gaze and stimulus position

Game performance measured children's response time (RT) and variance (RT-var) and the percentages of their positive and negative responses towards the Go and NoGo stimuli. RT was the average value of the time difference (R_t: equation 5–2) between when the character appeared (t_a) and when the subject's pressed the space bar (t_r); its standard deviation was RT-var. This study combined both of them as some children had a small percentage NoGo-negative that resulted in unreliable RT-var.

$$R_t = t_a - t_r \tag{5–2}$$

Absolute gaze position features estimated the overall subjects' gaze modulation during the experiment. Gaze trajectory area was calculated with Convex Hull algorithm; its value ranged from 0 to 1. Gaze velocity (equation 5–3) and acceleration (equation 5–4) along the x and y axes were computed as smoothed first and second time-derivatives of the corresponding coordinate locations, respectively. Gaze distance (equation 5–5) and angle (equation 5–6) were the Euclidian distance and the angle between two consecutive gaze positions: $g(t)$ and $g(t+1)$.

$$v = \sqrt{\partial_t x^2 + \partial_t y^2} \tag{5–3}$$

$$\partial v = \sqrt{\partial_t^2 x^2 + \partial_t^2 y^2} \tag{5–4}$$

$$d(t) = |g(t+1) - g(t)|_2 \tag{5–5}$$

$$a(t) = \arccos\left(\frac{g(t+1) \cdot g(t)}{|g(t+1)| \cdot |g(t)|}\right) \tag{5–6}$$

Gaze-to-object features measured the relative positions of participants' gaze and stimulus positions. Fixation time estimated the time difference between when the subject's gaze entered and when it left a stimulus' area; the area was a circle of radius 0.25 (measured by Euclidean distance) from the center of the

stimulus. Gaze-to-object difference was subjects' gaze positions minus object positions when the latter appeared on the screen. The probability and spectral densities of the difference were computed with kernel density estimation and Welch's method [84] ($nperseg = 32$), respectively. The probability distribution of the gaze-to-object difference was computed using 50 cells along each of the x and y axes.

Shannon Entropy was employed to estimated the irregularity of gaze distribution and was expressed as:

$$H_s = -\int p(x,y)\log_2(p(x,y))dxdy \tag{5–7}$$

where $p(x,y)$ was the probability density of a state in a two-coordinate plane; while for power spectral density, $p(x,y)$ was the sum of squared magnitudes of the Fourier transforms of the respective x and y components. Greater entropy shows more chaotic gaze modulation and suggests greater gaze dispersion. The final value of gaze entropy was normalized by diving by the maximum possible entropy $\log_2(N)$, in which N was the total number of states; in this study, N equaled to the total number of cells along the x and y axes.

This study estimated the temporal irregularity of gaze movement with Sample entropy (sen). It is equal to the negative natural logarithm of the probability that two subsequences of equal length m that are similar will remain similar at the next time step [85]; lower sample entropy means higher predictability within the original sequence. This study set m to two (low value of m can capture local irregularity) and estimated the distance between two template vectors with Chebyshev distance.

B Gaze-adjustment Features [4]

Gaze-adjustment features were Euclidean distance between the subject's gaze and stimulus position when the stimulus appeared on the screen. The value of the distance ranged from 0 to $\sqrt{2}$.

Since the appearance time of a stimulus depended on the subjects' RT, which

varied, each gaze-adjustment was represented by auto-regressive parameters (equation (5–8)). The model's lag L was set to two (average of AIC: -9.129); hence, each gaze-adjustment was represented by three variables: α, θ_2, and θ_1. The auto-regressive model was trained within 200 iterations.

$$y_t = \alpha + \sum_{i=1}^{L} \theta_i y_{t-i} \tag{5–8}$$

During extrapolation experiment, this study computed the average value of each coefficient for typical and ASD groups using the arithmetic means over the respective groups.

5.4.2 Method

A Participants

Before participating in the experiment, informed consent was got from teachers and parents on behalf of the children. The study was approved by the Research Ethics Committee of the Prefectural University of Hiroshima (letter no: 15MH070) and was conducted under the amended Declaration of Helsinki.

We recruited 35 typical children (24 male and 11 female) and 22 children with ASD symptoms (16 male and 6 female) from two local schools in Japan (Table 5.5). All children with ASD symptoms attended special schools and had been diagnosed by clinicians; 10 (+ one suspected) ASD children also had attention deficit symptoms, and seven of them were identified as having hyperactivity as well. All participants did not have any physical disabilities and their IQ scores were similar ($p > 0.913$). This study excluded one ASD child (male) and four typical children (1 male and 3 female) because their data were corrupted. Hence, only data of 21 ASD and 31 typical children were proceeded.

Before participating in the experiment, an instructor explained the game and its rules to teachers and children; the instructor asked the subjects to respond immediately when the stimulus appeared. Then, the instructor calibrated the eye-tracker for each subject.

Table 5.5: Differences for age and Development Quotient (DQ) scores were insignificant ($p > 0.05$). The average and standard deviation (STD) of age and DQ score of typical and ASD groups. All children participated in this study were Japanese.

	Male/Female	Age (Mean ± S.D.)	DQ (Mean ± S.D.)
Typical	24/11	5.0 ± 0.6	96.1 ± 3.0
ASD	16/6	4.6 ± 0.4	95.7 ± 10.4
Student t-test	-	0.234	0.913

This study utilized a notebook equipped with an eye-tracker and web camera. The participants were seated in front of the notebook. They responded to the stimulus by pressing the spacebar on the keyboard (Fig. 5.9). The proportion of the Go and NoGo stimuli was uniform and the order of appearance was set in advance; their appearance time was 700 ms; the minimum and maximum of the waiting period were 700 and 1000 ms.

All children took one-minute training before participating in four-minute evaluation. The instructor assisted the subjects during the training.

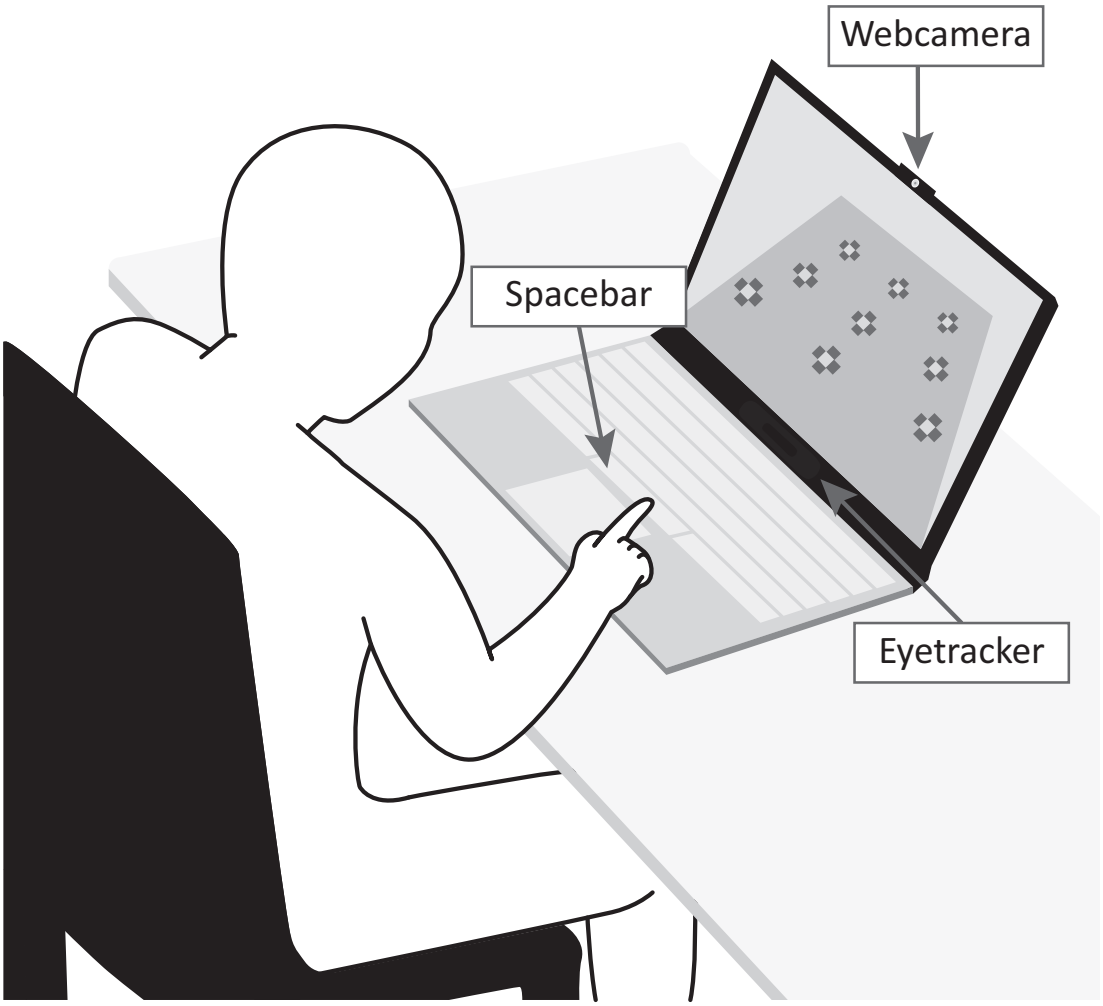


Fig. 5.9: Experimental protocol of this study. The distance between the child and the monitor was about 60 cm. The notebook was equipped with a web camera and an eye tracker.

B Preprocessing

Figure 5.10 shows the pipeline for extracting spatial and gaze-adjustment features from response and gaze data. Before extracting the features, preprocessing was performed to eliminate noise and redundant data.

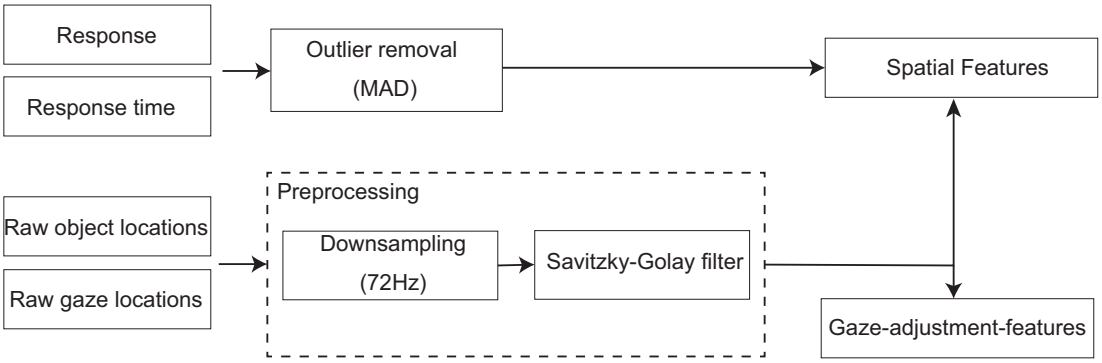


Fig. 5.10: Features extraction pipeline used in this study. The inputs consists of gaze and object locations, response, and response time.

The responses whose RT was less than a threshold were considered as outliers; 6.6% of typical and 7.3% of ASD data were removed. The threshold was the RT's median absolute deviation [86] (104.75 ms) multiplied by a constant scale factor of the normal distribution (1.4826): $1.4826 \times 104.75 = 155.30$ ms. Redundancy in gaze data was minimized by down-sampling them from from 144 Hz to 72 Hz; a Savitzky-Golay filter [87] ($n = 5$ and $poly = 2$) was used to perform smoothing to prevent artifacts during numerical differentiation.

5.4.3 Statistical Analysis

Statistical comparison between groups was performed with Student t and Mann-Whitney U [88] tests. Multiple comparisons were conducted with ANOVA test. This study assumed that data distributions of typical and ASD groups were independent. The false discovery rate associated with multiple comparisons was controlled using the BenjaminiHochberg procedure [89], at the level 0.05. Hence, significant difference was decided based on the critical value obtained from the

BenjaminiHochberg procedure.

The effect size of a variable was calculated using Cohen's d [90]; the first and second groups were typical and ASD populations, respectively.

5.4.4 Results

A Spatial Features

A Statistical comparison indicated a significant difference (by both Student t and Mann-Whitney U tests) between typical and ASD groups in eight features: variance of fixation time, average and entropy of gaze acceleration, spectral entropy of gaze-to-object-distance, sample entropy of gaze distance, gaze angle, gaze-to-obj-distance, and velocity. Also, a substantial difference between the groups was observed in the percentage of Go-negative and variance of gaze acceleration. For all significantly different variables, the mean values of the typical group was smaller than the ASD group. The effect sizes of the corresponding variables were large ($|d| > 0.8$), except that of the average gaze acceleration, which was moderate ($d = -0.763$). Similarly, the effect size of substantially different variable was moderate ($|d| > 0.5$), which suggested considerably different to those variables' mean value between the groups.

Contrary to gaze-related features, insignificant difference was observed in the average response time and its variance. Within-group mean values, however, indicated that ASD children's response was higher than typical children ($d = -0.518$).

Similar results were observed in the statistical results between typical and ASD subjects without ADHD. Statistical analysis demonstrated a significant difference for those eight variables, except spectral entropy and entropy of gaze-to-object-distance, and an average of gaze acceleration. Besides, both Student t and Mann-Whitney U tests suggested that percentage of Go positive and negative of typical subjects differed significantly from those of ASD participants without ADHD symptoms. The effect sizes of those variables were large ($|d| > 0.8$) and had positive signs except that of the Go-positive.

A comparison between typical and ASD children with ADHD yielded different results.

Table 5.6: Statistical significance between-group (p) differences of individual spatial features, as measured by Student t, Mann-Whitney U, and ANOVA tests. TD stands for typical. + and − mean with and without, respectively.* indicates significant p-value after controlling false discovery rate at level 0:05.

	TP vs ASD		TP vs ASD-ADHD		TP vs ASD+ADHD		ASD-ADHD vs ASD+ADHD		ANOVA (p)
	Student (p)	Whitney (p)	Student (p)	Whitney (p)	Student (p)	Whitney (p)	Student (p)	Whitney (p)	
Go-positive	0.042	0.040	***0.007**	***0.005**	0.483	0.458	0.216	0.085	0.099
Go-negative	0.039	0.033	***0.006**	***0.005**	0.501	0.392	0.186	0.074	0.096
NoGo-positive	0.712	0.404	0.585	0.313	0.948	0.458	0.763	0.275	0.647
NoGo-negative	0.797	0.411	0.818	0.376	0.856	0.494	0.984	0.376	0.596
RT	0.595	0.334	0.259	0.284	0.690	0.470	0.323	0.349	0.857
RT-var	0.073	0.032	0.028	***0.012**	0.547	0.277	0.200	0.123	0.197
Trajectory-area	0.845	0.321	0.580	0.344	0.837	0.375	0.626	0.349	0.124
Velocity-avg	0.066	0.102	0.227	0.139	0.036	0.185	0.352	0.486	***0.006**
Velocity-var	0.042	0.070	0.257	0.120	0.018	0.127	0.261	0.458	***0.007**
Acceleration-avg	***0.009**	***0.002**	0.031	***0.008**	***0.007**	0.020	0.309	0.486	***0.001**
Acceleration-var	0.022	0.035	0.121	0.058	***0.012**	0.104	0.288	0.486	***0.003**
Fixation-avg	0.422	0.404	0.547	0.195	0.093	0.093	0.113	0.096	0.254
Fixation-var	***0.001**	***0.001**	***0.016**	***0.010**	***0.004**	***0.002**	0.639	0.275	***0.004**
Distance-avg	0.117	0.234	0.395	0.237	0.062	0.341	0.370	0.458	***0.013**
Distance-var	0.095	0.136	0.503	0.246	0.036	0.147	0.261	0.349	***0.025**
Angle-avg	0.292	0.500	0.791	0.454	0.159	0.446	0.405	0.458	0.034
Angle-var	0.241	0.341	0.978	0.500	0.076	0.247	0.244	0.299	0.062
Distance-sen	***0.001**	***0.001**	***0.001**	***0.006**	***0.004**	***0.011**	0.765	0.403	***0.001**
Angle-sen	***0.001**	***0.001**	***0.001**	***0.004**	***0.006**	0.019	0.760	0.275	***0.001**
Velocity-sen	***0.000**	***0.000**	***0.001**	***0.004**	***0.001**	***0.002**	0.718	0.430	***0.000**
Spatial-en	0.993	0.195	0.845	0.115	0.873	0.458	0.845	0.230	0.223
Gaze-obj-en	***0.004**	***0.004**	0.030	0.031	***0.007**	***0.011**	0.552	0.349	***0.006**
Gaze-obj-sen	***0.001**	***0.001**	***0.010**	***0.016**	***0.002**	***0.002**	0.566	0.458	***0.001**
Gaze-obj-spe	***0.003**	***0.002**	0.023	***0.010**	0.021	0.025	0.772	0.376	***0.013**

Table 5.7: d denotes Cohen's effect size measure of spatial features between groups. TP vs ASD means that typical (TP) and ASD population were treated as the first and second group, respectively. + and − respectively mean with and without.

	TP x ASD d	TP x ASD-ADHD d	TP x ASD+ADHD d	ASD-ADHD x ASD+ADHD d
Go-positive	0.591	0.998	0.258	-0.559
Go-negative	-0.598	-1.014	-0.247	0.600
NoGo-positive	0.105	0.193	0.024	-0.134
NoGo-negative	-0.073	-0.081	-0.066	0.009
RT	0.151	0.402	-0.146	-0.443
RT-var	-0.518	-0.800	-0.221	0.581
Trajectory-area	0.056	0.196	-0.075	-0.216
Velocity-avg	-0.531	-0.431	-0.791	-0.417
Velocity-var	-0.589	-0.404	-0.902	-0.506
Acceleration-avg	-0.763	-0.783	-1.042	-0.457
Acceleration-var	-0.668	-0.556	-0.964	-0.478
Fixation-avg	0.229	-0.213	0.626	0.726
Fixation-var	-0.997	-0.883	-1.119	-0.208
Distance-avg	-0.451	-0.302	-0.699	-0.401
Distance-var	-0.482	-0.237	-0.789	-0.506
Angle-avg	-0.301	-0.094	-0.522	-0.372
Angle-var	-0.335	-0.010	-0.662	-0.525
Distance-sen	-1.026	-1.218	-1.129	0.133
Angle-sen	-1.020	-1.280	-1.054	0.136
Velocity-sen	-1.133	-1.307	-1.267	0.160
Spatial-en	-0.002	-0.069	0.059	0.087
Gaze-obj-en	-0.864	-0.787	-1.035	-0.265
Gaze-obj-sen	-0.973	-0.953	-1.184	-0.255
Gaze-obj-spe	-0.880	-0.833	-0.875	-0.128

Table 5.8: The average and standard deviation (STD) values of each spatial features for typical (TP), ASD, ASD without ADHD, and ASD with ADHD groups. Percentage scale is used to express Go positive and negative, and NoGo positive and negative. While, RT, RT-var, fixation-avg, and fixation-var are expressed in millisecond.

	TP (Mean ± S.D.)	ASD (Mean ± S.D.)	ASD-ADHD (Mean ± S.D.)	ASD+ADHD (Mean ± S.D.)
Go-positive	30.8 ± 9.9%	24.2 ± 12.9%	20.8 ± 10.5%	27.9 ± 14.8%
Go-negative	19.0 ± 10.0%	25.7 ± 12.9%	29.3 ± 10.7%	21.7 ± 14.4%
NoGo-positive	46.2 ± 2.7%	45.9 ± 3.4%	45.7 ± 3.0%	46.1 ± 4.0%
NoGo-negative	4.0 ± 2.8%	4.2 ± 3.0%	4.2 ± 2.7%	4.2 ± 3.5%
RT	545 ± 34.3	539 ± 48.7	529 ± 56.2	551 ± 38.6
RT-var	117 ± 36.9	136 ± 35.7	146 ± 31.7	126 ± 38.3
Trajectory-area	0.521 ± 0.103	0.515 ± 0.134	0.501 ± 0.109	0.530 ± 0.161
Velocity-avg	0.0088 ± 0.0010	0.0098 ± 0.0024	0.0093 ± 0.0011	0.0103 ± 0.0034
Velocity-var	0.0176 ± 0.0024	0.0198 ± 0.0051	0.0186 ± 0.0025	0.0212 ± 0.0069
Acceleration-avg	0.0035 ± 0.0006	0.0044 ± 0.0016	0.0040 ± 0.0007	0.0047 ± 0.0022
Acceleration-var	0.0076 ± 0.0015	0.0092 ± 0.0034	0.0085 ± 0.0017	0.0101 ± 0.0046
Fixation-avg	405 ± 20.6	399 ± 28.9	409 ± 12.6	389 ± 38.0
Fixation-var	174 ± 8.2	183 ± 8.3	182 ± 8.9	183 ± 8.1
Distance-avg	0.0160 ± 0.0015	0.0171 ± 0.0035	0.0164 ± 0.0016	0.0178 ± 0.0049
Distance-var	0.0293 ± 0.0033	0.0316 ± 0.0063	0.0301 ± 0.0031	0.0333 ± 0.0084
Angle-avg	0.0130 ± 0.0013	0.0136 ± 0.0026	0.0131 ± 0.0011	0.0141 ± 0.0036
Angle-var	0.0285 ± 0.0039	0.0304 ± 0.0073	0.0285 ± 0.0036	0.0323 ± 0.0098
Distance-sen	0.147 ± 0.031	0.202 ± 0.076	0.207 ± 0.083	0.196 ± 0.072
Angle-sen	0.124 ± 0.026	0.169 ± 0.063	0.173 ± 0.063	0.165 ± 0.065
Velocity-sen	0.137 ± 0.034	0.200 ± 0.077	0.206 ± 0.088	0.193 ± 0.069
Spatial-en	0.894 ± 0.011	0.894 ± 0.020	0.895 ± 0.016	0.893 ± 0.024
Gaze-obj-en	0.484 ± 0.013	0.497 ± 0.017	0.494 ± 0.015	0.499 ± 0.020
Gaze-obj-sen	0.130 ± 0.018	0.153 ± 0.031	0.150 ± 0.027	0.158 ± 0.036
Gaze-obj-spe	0.354 ± 0.008	0.361 ± 0.006	0.360 ± 0.005	0.361 ± 0.008

B Gaze-adjustment Results

The Mann-Whitney U test showed a significant difference between the groups in the mean values of α, θ_1 and θ_2. The Student t-test and effect size ($|d| < 0.2$), however, suggested that typical children and ASD children's gaze-adjustment did not differ.

Separating gaze-adjustment features according to response types (Go-positive, Go-negative, NoGo-positive, and NoGo-negative) yielded statistically significant differences ($n = 52$, $p < 0.023$) between typical and ASD children in all autoregressive coefficients by the Mann-Whitney U test, as well as greater effect size (mean $|d| > 0.4$). Also, the student t-test demonstrated a significant difference between the groups in gaze-adjustment-features of Go-negative and NoGo-positive. In brief, the results signified ASD gaze modulation differed from the typical when they responded incorrectly to the Go stimulus and correctly to the NoGo stimulus.

Extrapolation of the gaze-to-obj distance in time using the average values of the autoregressive coefficients suggests that separating the features (Fig. 5.11C - J) resulted in a more apparent difference between the groups than mixing them (Fig. 5.11A, B). ASD children adjusted their gaze to the stimulus position slower when they responded correctly to the Go and NoGo characters and when they reacted incorrectly to the latter stimulus (Fig. 5.11C, G, I); the velocity of their extrapolated gaze-adjustment (Fig. 5.11D, H, J) was ± 0.0014 slower compared to the typical children (the velocity of extrapolated gaze-adjustment was computed by averaging the negative of the first derivative of the extrapolated gaze-to-obj distance over time). However, ASD children modulated their gaze similarly to the typical subjects when they missed the Go stimulus (Fig. 5.11E, F).

Table 5.9: Statistical significance between-group (p) differences of individual gaze-adjustment features, as measured by Student t, Mann-Whitney U, and ANOVA tests. + and − mean with and without, respectively. * indicates significant p-value after controlling false discovery rate at level 0:05.

	TP vs ASD		TP vs ASD - AD		TP vs ASD + ADHD		ASD - ADHD vs ASD + ADHD		ANOVA (p)
	Student (p)	Whitney (p)	Student (p)	Whitney (p)	Student (p)	Whitney (p)	Student (p)	Whitney (p)	
Average									
C1	0.646	***0.000**	0.750	***0.000**	0.671	***0.000**	0.956	***0.000**	0.899
C2	0.282	***0.023**	0.262	***0.000**	0.594	***0.023**	0.620	***0.023**	0.506
C3	0.272	***0.000**	0.205	***0.000**	0.660	***0.000**	0.509	***0.000**	0.448
Go-positive									
C1	0.180	***0.000**	0.418	***0.000**	0.212	***0.000**	0.613	***0.000**	0.370
C2	0.185	***0.023**	***0.004**	***0.000**	0.862	***0.023**	0.356	0.023	0.164
C3	0.391	***0.000**	***0.004**	***0.000**	0.778	***0.000**	0.317	***0.000**	0.213
Go-negative									
C1	0.206	***0.000**	0.068	***0.000**	0.654	***0.000**	0.516	***0.000**	0.295
C2	0.016	***0.000**	0.106	***0.023**	***0.017**	***0.000**	0.433	***0.000**	0.036
C3	***0.006**	***0.000**	0.041	***0.000**	***0.011**	***0.000**	0.509	***0.000**	0.018
NoGo-positive									
C1	0.135	***0.000**	0.344	***0.000**	0.062	***0.000**	0.383	***0.000**	0.130
C2	***0.000**	***0.023**	***0.001**	***0.023**	***0.000**	***0.023**	0.378	***0.023**	***0.000**
C3	***0.000**	***0.000**	***0.000**	***0.000**	***0.000**	***0.000**	0.399	***0.000**	***0.000**
NoGo-negative									
C1	0.877	***0.000**	0.838	***0.000**	0.586	***0.000**	0.587	***0.000**	0.834
C2	0.688	***0.000**	0.984	***0.000**	0.524	***0.000**	0.493	***0.023**	0.781
C3	0.794	***0.000**	0.890	***0.000**	0.582	***0.000**	0.498	***0.000**	0.811

Table 5.10: d denotes Cohen's effect size measure of gaze-adjustment features between groups. TP vs ASD means that typical and ASD population were treated as the first and second group, respectively. + and − respectively mean with and without.

	TP vs ASD (d)	TP vs ASD - ADHD (d)	TP vs ASD + ADHD (d)	ASD - ADHD vs ASD + ADHD (d)
Average				
C1	-0.065	-0.056	-0.077	-0.012
C2	0.153	0.198	0.097	-0.109
C3	-0.156	-0.223	-0.080	0.145
Go-positive				
C1	-0.385	-0.287	-0.462	-0.225
C2	0.379	1.061	0.064	-0.414
C3	-0.245	-1.081	0.103	0.449
Go-negative				
C1	0.362	0.659	0.164	-0.289
C2	0.703	0.581	0.909	0.350
C3	-0.807	-0.740	-0.974	-0.294
NoGo-positive				
C1	-0.429	-0.336	-0.698	-0.390
C2	1.155	1.246	1.493	0.394
C3	-1.259	-1.340	-1.589	-0.377
NoGo-negative				
C1	0.044	-0.072	0.199	0.241
C2	-0.114	0.007	-0.234	-0.306
C3	0.074	-0.049	0.202	0.302

Table 5.11: The average and standard deviation (STD) values of the gaze-adjustment features for typical (TD), ASD, ASD without ADHD, and ASD with ADHD groups. α, θ_1, and θ_2 are the auto-regressive model's constant term, first, and second coefficients, respectively.

	Typical (Mean ± S.D.)	ASD (Mean ± S.D.)	ASD without ADHD (Mean ± S.D.)	ASD with ADHD (Mean ± S.D.)
Average				
C1	0.004 ± 0.012	0.005 ± 0.013	0.005 ± 0.015	0.005 ± 0.010
C2	1.488 ± 0.221	1.456 ± 0.189	1.446 ± 0.179	1.467 ± 0.199
C3	-0.520 ± 0.206	-0.489 ± 0.191	-0.475 ± 0.176	-0.503 ± 0.206
Go-positive				
C1	0.002 ± 0.009	0.005 ± 0.007	0.004 ± 0.006	0.006 ± 0.008
C2	1.533 ± 0.060	1.488 ± 0.167	1.455 ± 0.098	1.525 ± 0.213
C3	-0.561 ± 0.054	-0.529 ± 0.193	-0.487 ± 0.095	-0.575 ± 0.254
Go-negative				
C1	0.003 ± 0.002	0.002 ± 0.005	0.001 ± 0.004	0.003 ± 0.005
C2	1.572 ± 0.059	1.521 ± 0.084	1.536 ± 0.063	1.505 ± 0.099
C3	-0.601 ± 0.054	-0.549 ± 0.077	-0.560 ± 0.058	-0.536 ± 0.092
NoGo-positive				
C1	0.003 ± 0.002	0.005 ± 0.008	0.003 ± 0.003	0.007 ± 0.011
C2	1.606 ± 0.031	1.540 ± 0.081	1.555 ± 0.059	1.523 ± 0.097
C3	-0.631 ± 0.028	-0.571 ± 0.066	-0.583 ± 0.050	-0.558 ± 0.077
NoGo-negative				
C1	0.010 ± 0.021	0.009 ± 0.023	0.011 ± 0.029	0.006 ± 0.012
C2	1.241 ± 0.321	1.274 ± 0.238	1.238 ± 0.218	1.314 ± 0.252
C3	-0.286 ± 0.295	-0.306 ± 0.230	-0.272 ± 0.219	-0.344 ± 0.236

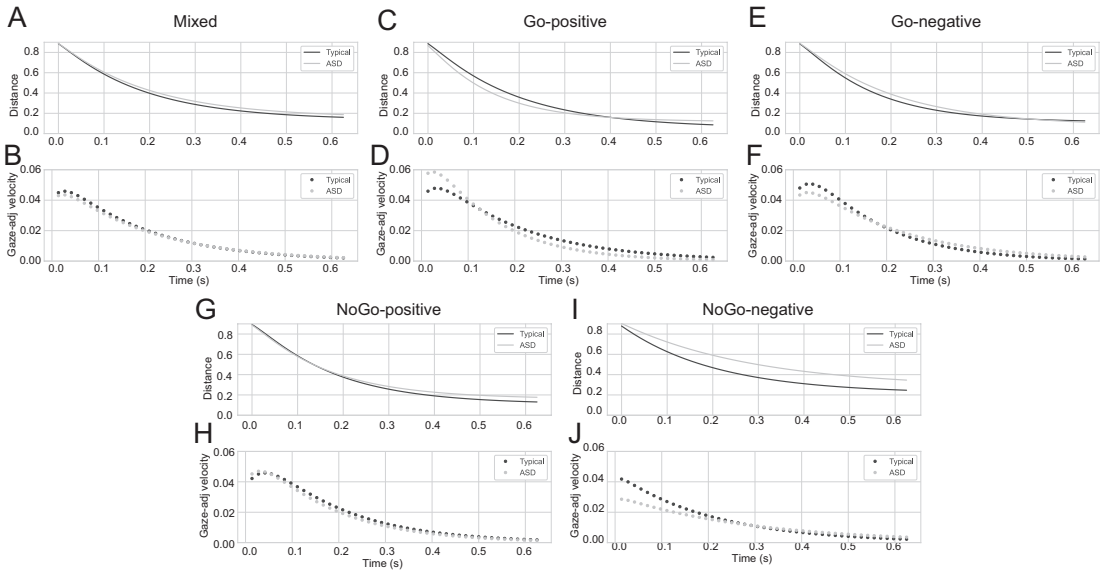


Fig. 5.11: Extrapolating results of Auto-regressive model using the average of parameters. Gaze extrapolation results using mixed (A), Go positive (C) and negative (E), and NoGo positive (G) and negative (I) coefficients. (B, D, F, H, J) show respectively the extrapolated gaze-to-obj distance and velocity results for mixed (typical-avg: 0.0161, ASD-avg: 0.0156), Go positive (typical-avg: 0.0178, ASD-avg: 0.0165) and negative (typical-avg: 0.0169, ASD-avg: 0.0173), and NoGo positive (typical-avg: 0.0170, ASD-avg: 0.0158) and negative (typical-avg: 0.0141, ASD-avg: 0.0124) coefficients. Solid and dotted green lines represent, respectively, typical children's extrapolated gaze-to-obj distance and the negative of its first derivative (gaze-adjustment velocity) over time. ASD children's extrapolated gaze-to-obj distance and gaze-adjustment velocity are represented by sold and dotted orange lines, respectively.

A significant difference was observed between the typical and ASD children without ADHD when they responded correctly towards the Go and NoGo stimuli ($n = 52$, $p <= 0.004$); the corresponding effect size of those variables were large ($|d| > 0.8$).

Comparison results of typical children to ASD children with ADHD indicated that the former responded differently from the latter during Go-negative and

NoGo-positive ($n = 52$, $p <= 0.017$). The results also demonstrated that ASD children with and without ADHD did not differ significantly.

The ANOVA test showed a significant difference ($n = 52$, $p < 0.04$) among those three groups for gaze-adjustment features when the subjects responded correctly to the NoGo stimulus; the difference was insignificant in the other conditions.

The extrapolation results of ASD children with and without ADHD symptoms (Fig. 5.12 A - J) suggested that the former tended to adjust their gaze to the stimulus position slightly faster than the latter, with respective extrapolation gaze-adjustment velocities of 0.0153 and 0.0156. Both groups had lower gaze-modulation speed compared to typical participants, whose average velocity was 0.0164.

5.5 Concluding Remarks

Relation between Response and Gaze Behavior during the Go/NoGo Task

As we expected, we found that subjects' response related to their gaze modulation. Subjects with bigger gaze trajectory tended to respond slower with higher variance. The results suggested that this relationship was stronger among male than female subjects.

Second, clustering using K-means resulted in two separable clusters. Subjects belong to different clusters had different gaze patterns: the ones that focused their gaze on the middle of the screen, and the ones that adjusted their gaze to stimulus position. Also, their RT and RT-var differed significantly, in which the members of the first cluster responded faster with lower variance.

Finally, the preliminary result of a child with ASD symptoms showed that gaze modulation of ASD and typical children differed. Typical children tended to adjust their gaze position to stimulus position, while the child with ASD seemed to have difficulty to modulate his gaze. These findings agreed with the previous

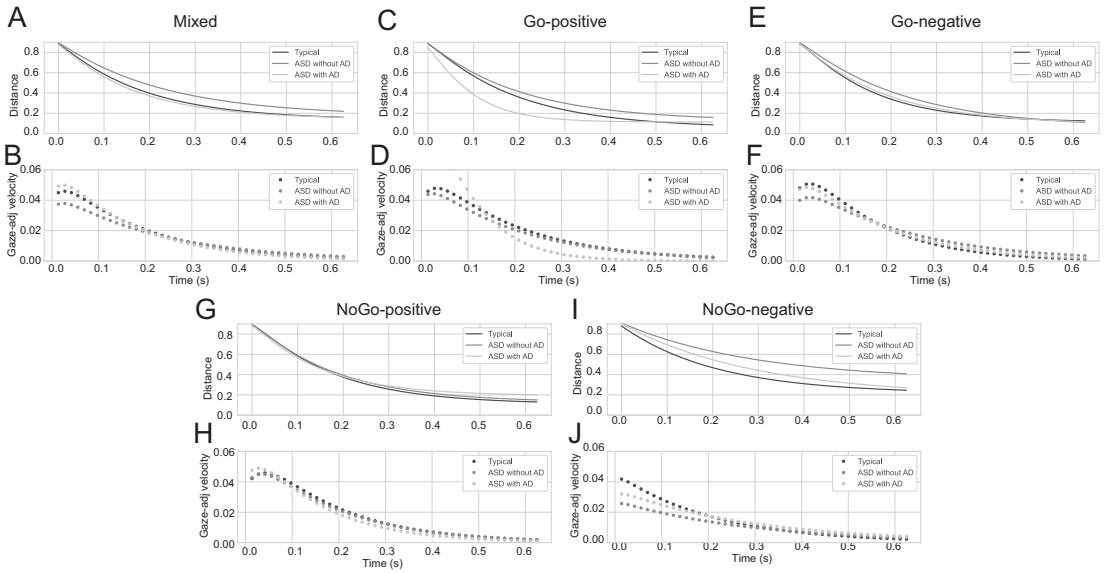


Fig. 5.12: Extrapolating results of Auto-regressive model using the average of parameters. Gaze extrapolation results using mixed (A), Go positive (C) and negative (E), and NoGo positive (G) and negative (I) coefficients. (B, D, F, H, J) show respectively the extrapolated gaze-to-obj distance and velocity results for mixed (typical-avg: 0.0161, ASD without ADHD-avg: 0.0150, ASD with ADHD-avg: 0.0160), Go positive (typical-avg: 0.0178, ASD without ADHD-avg: 0.0162, ASD with ADHD-avg: 0.0162) and negative (typical-avg: 0.0169, ASD without ADHD-avg: 0.0175, ASD with ADHD-avg: 0.0170), and NoGo positive (typical-avg: 0.0170, ASD without ADHD-avg: 0.0165, ASD with ADHD-avg: 0.0152) and negative (typical-avg: 0.0141, ASD without ADHD-avg: 0.0111, ASD with ADHD-avg: 0.0140) coefficients. Solid and dotted green lines represent, respectively, typical children's extrapolated gaze-to-obj distance and the negative of its first derivative (gaze-adjustment velocity). Extrapolated gaze-to-obj distance and gaze-adjustment velocity of ASD children with and without ADHD symptoms are represented by purple and pink colors, respectively.

works [13, 91] that observed a greater irregularity of gaze movement in children with ASD symptoms during face-to-face conversation.

Response and Gaze Behavior of Children with ASD Symptoms

This chapter investigates the relationship between children's response and gaze behavior with ASD symptoms. The experimental results involving 35 typical and 22 ASD subjects suggested a significant difference between typical and ASD children. The results suggested that children with ASD symptoms had lower accuracy and greater randomness in visual tracking of the stimuli: the relative gaze-to-object difference was less steady over time than for typical subjects, and predictability of ASD subjects' gaze was lower than measured by sample entropy of both angle and distance. A greater irregularity of gaze distance and angle may show that ASD children over-interpreted the information of a stimulus, causing more unintentional viewing behavior [92]. The higher value of ASD children's gaze-to-object entropy suggested less structured tracking in a spatial sense, while a greater value of sample entropy value demonstrated lowered predictability of the gaze-to-object difference as a function of time. Likewise, greater spectral entropy signified less structure of the frequency content of ASD subjects' gaze signals.

Last, the statistical comparisons suggested that game performance of typical and ASD subjects without ADHD symptoms differed substantially as indicated by percentage of Go positive and negative, and response time variance. Also, the results signified that different gaze modulation between typical and ASD subjects with ADHD symptoms was more pronounced than the ones without ADHD symptoms.

Chapter 6

Diagnostic Support System Using Interpretable Deep Distance Learning to Identify Developmental Disorder Symptoms in Children

6.1 Introduction

A diagnostic support system requires interpretable and evidence-based estimation results [18]. Explainable machine learning, e.g. Linear regression and decision tree, offer prediction results that stockholders can easy to understand. The performance of such algorithm, however, are outperformed by more sophisticated algorithms such as Deep Learning Neural networks (DNN) that provide a high accuracy rate in many real-world classification problems.

One of DNN's drawbacks is black-box structure, in which studying structure of it does not give insight into what it has learned from the data. Recent studies have attempted to solve that issue by proposing algorithms to interpret the prediction

results of DNN.

Model-Agnostic methods interpret black-box models' predictions by describing how features affect the average prediction results (global methods) or individual predictions' (local methods) [93]. Global methods can understand the general mechanism of the model. While local methods are useful to understand why the model's estimation result of an input.

This study uses one of Local methods, SHAP values [30], to interpret the model's individual predictions to the users. The primary goal is to explain to the users why features from a subject is much closer to one group than the others.

Also, as comorbidities are common among children with disorder symptoms, we assume that data of disorder children have many classes that overlap each other. Employing classification DNN requires us to provide great number of data points per class. Therefore, this study uses deep distance learning (DDL) [94] to measure similarity between a query in existing support vector groups.

Employing DDL gives two advantages to our proposed system. First, the proposed system can learn to new classes with only few samples. Second, the proposed system can perform retrievals to fetch support vectors that are similar to the query, providing evidence-based results. Furthermore, combining it with SHAP values [30], the proposed system can produce interepretable estimation results based on the feature's contribution(Fig. 6.1).

6.2 Proposed System

To deliver evidence-based results, we employed Deep Distance Learning to perform similarity measurement and retrieval. Then, the measurement results were interpreted by employing SHAP value that estimated how much each feature contributed to the results (Fig. 6.2).

This study presents a novel training loss for Deep Distance Learning that combines hard-triplet with prototypical network's loss. The proposed loss, named Cluster Hard Triplet loss (CsTL), compared an anchor with positive and hard

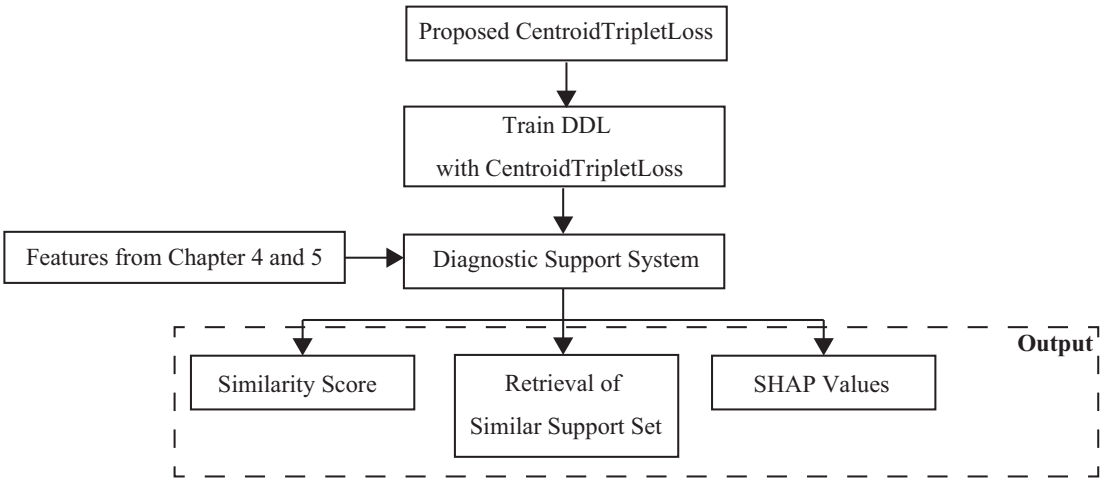


Fig. 6.1: Study flow diagram of Chapter 6.

negative centroids. Positive centroid was measured as the average embedding vectors that had the same labels as the anchor. While hard negative centroid was the closest centroid that had different from the anchor.

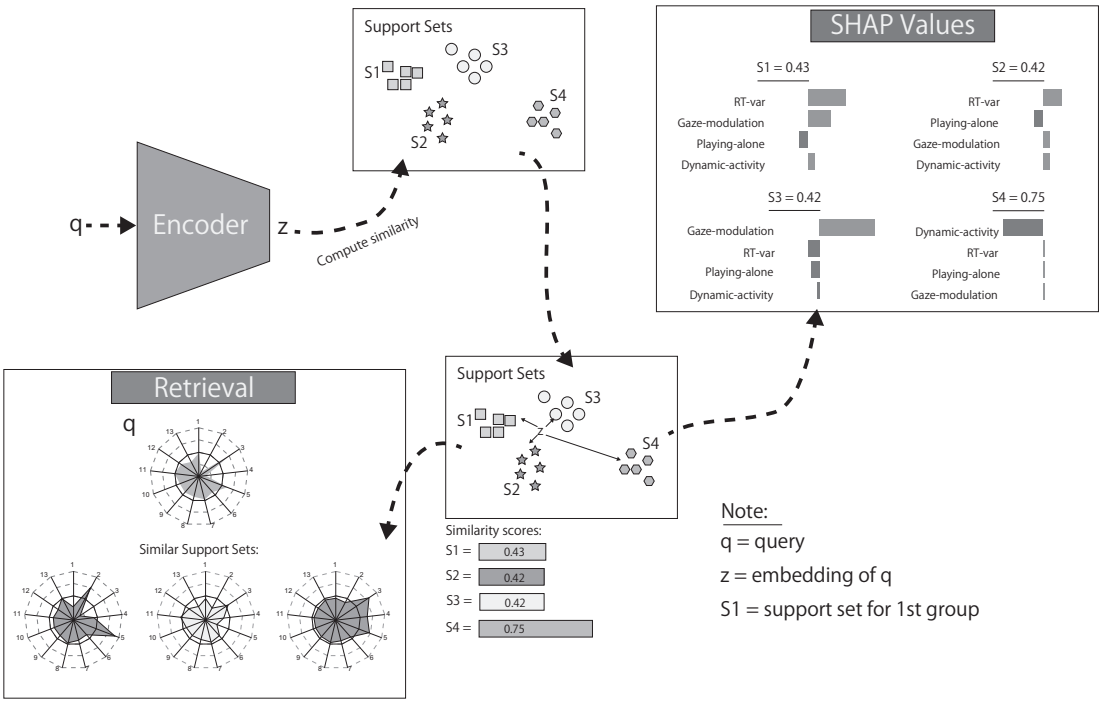


Fig. 6.2: A decision support system employing Deep Distance Learning and SHAP values.

6.2.1 Proposed Loss: Cluster Hard Triplet Loss

CsTL was motivated by Dunn index, that aimed to identify compactness and separability of clusters. CsTL worked by minimizing and maximizing intraclass and interclass variance, respectively. Minimizing the loss is like maximizing the Dunn index that results in compact and separable embedding vectors (Eq. 6–1).

Similar to the triplet-loss, CsTL did not force the distance between the anchor and the positive centroid into zero, which prevented class collapse. Also, comparing the anchor to centroids allowed the proposed loss to estimate mixture density in the data. The use of Euclidian distance in this study assumed that data-distribution was spherical.

$$\begin{aligned} \mathcal{L}(\theta; \mathcal{X}) &= \sum_{i=1}^{P} \sum_{a=1}^{K} [m + d(x_{a,i}, c_i) - \min_{\substack{n=1..K \\ n \neq i}} d(x_{a,i}, c_n)]_+ \\ d(x_a, x_b) &= D(f_\theta(x_a), f_\theta(x_b)) \end{aligned} \tag{6–1}$$

6.2.2 Interpretable Machine Learning

SHAP value [30] is the solution concept in cooperative game theory and can be classified as a local model-agnostic method. In a cooperative game, a coalition of players must cooperate to get certain games. SHAP value estimates how much a player contributes to the overall cooperation.

To interpret estimation results of a machine learning model, SHAP value treats the estimation task as the cooperative game. Then, it computes the contributions of coalition of players (features) by subtracting the actual prediction for the instance from the average predictions for all instances. Finally, SHAP value is estimated by averaging marginal contribution of a feature value across all coalitions.

6.3 Experiment

A Datasets

OmniGlot [**95**] comprises 1623 distinct characters from 50 different alphabets. The study followed the experiment protocol of previous studies [96–98] that used 1200 characters as training and 423 characters as testing. Data augmentation was performed by rotating each character by multiples of 90 degrees (90, 180, 270), totaling 4800 and 1692 classes for training and testing data, respectively. Before training, input images were resized to 28x28, and their pixel values were normalized to 0 to 1.

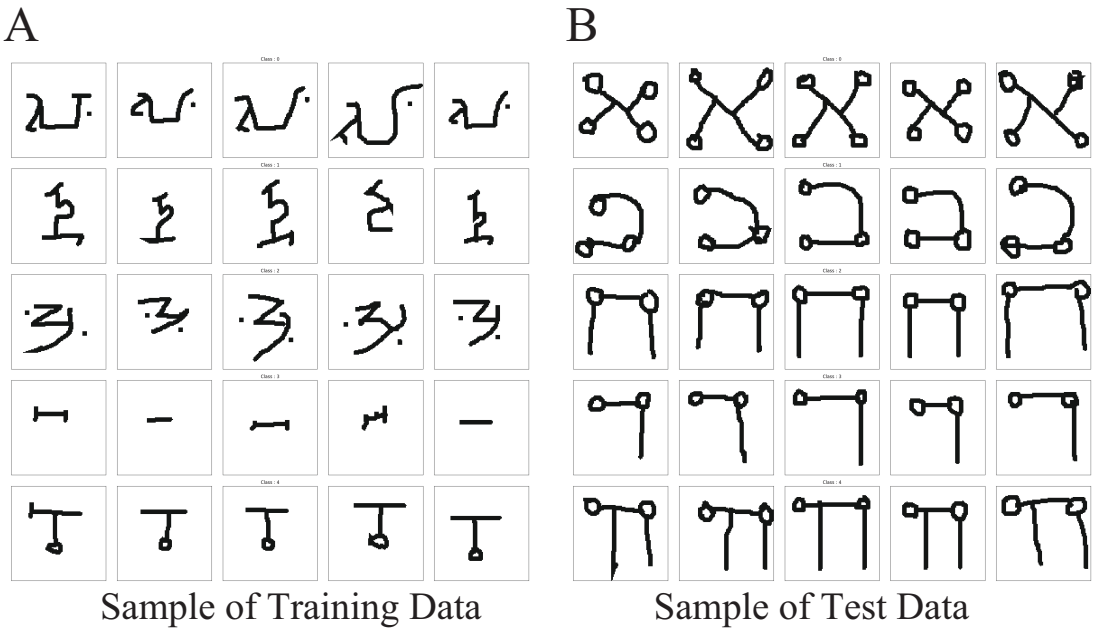


Fig. 6.3: Sample of training (A) and test (B) data of Omniglot dataset. The characters in test and training data differ.

6.3.1 Implementation Detail

Network architectures: On the Omniglot experiment, embeddings were extracted from input images using four convolutional blocks followed by a dense

network with 64-dimensional output space; the outputs of the network were normalized with $l2$-norm. The architecture of a convolutional block comprised a 64-filter 3×3 convolution, batch normalization layer, a ReLU activation function, and a 2×2 max-pooling layer.

Training procedures: The Omniglot network was trained using Adam with default hyperparameter values ($lr = 1e-3, \beta_1 = 0.9, \beta_2 = 0.999$). Exponential decay was employed to decrease the value of lr during the training; the decay steps and rate were 1000 and 0.8, respectively. Training was performed on 5000 episodes, in which in each episode P classes with 20-shot were used as training data. ResNet50 was trained using the same Optimizer with different values of learning rate ($lr = 5e-4$). Training was performed on 10000 episodes that comprised 5-shot for each of them.

6.3.2 Evaluation Protocol

In the classification problem, the model was evaluated with the N-way-K-shot protocol. Each episode comprised one query image and K support images for each class. The query and support images were different instances that were taken randomly. For each evaluation test, we fixed the value of the random seed.

6.4 Results

6.4.1 Ablation Study

Evaluated on Omniglot, experimental results demonstrated CsTL was insensitive to the number of classes in a training episode (Fig. 6.4). Although a higher value of P seemed to yield a higher accuracy rate, the improvement was statistically insignificant (avg of $p-\text{val} > 0.05$).

Comparison results between centroid and median indicated that using centroid as clusters' center yielded an average of 0.05% lower accuracy rate on Omniglot.

The results of using different values of m showed different outcomes on Om-

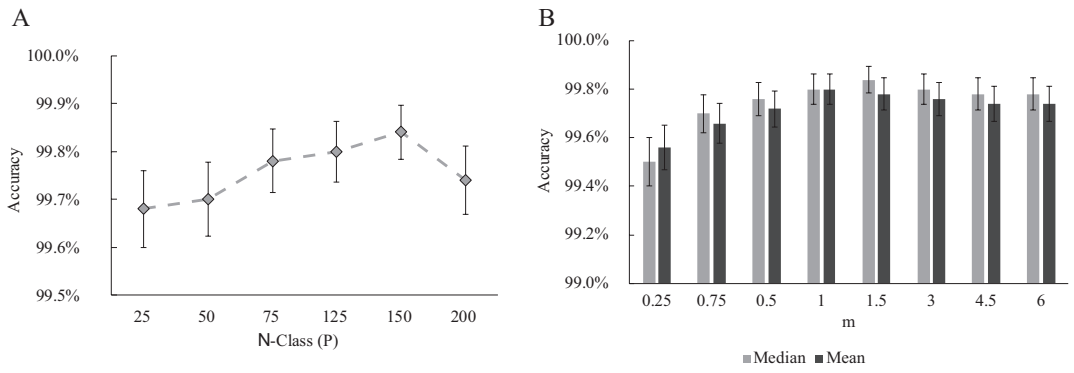


Fig. 6.4: (A) Average accuracy rate of 5-way-5-shot over 1000 episodes on Omniglot. (B) Average accuracy rate of 5-way-5-shot over 1000 episodes on Omniglot. The clusters' center was computed using median, and the value of m was set respectively to 1.5.

niglot and Market. Evaluated on Omniglot, different values of m did not lead to a significant discrepancy in accuracy rate.

6.4.2 Omniglot Results

Table 6.1 shows the Comparison results of CsTL with the state-of-the-art on Omniglot dataset. Compared to Matching networks and its variants [99–101], CsTL attained higher accuracy rate on all evaluation protocols. Yet, it achieved comparable and lower accuracy rate than Prototypical Net [98] and Model-Agnostic methods [102–105].

Table 6.1: Average accuracy rate (%) and standard-deviation (%) of CsTL and previous works on Omniglot dataset over 1000 test-episode.

		5-way		20-way		50-way		100-way	
Method	Fine Tune	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Memory Module	Y	98.40	99.60	95.00	98.60	-	-	-	-
MAML	Y	98.7 ± 0.4	99.9 ± 0.1	95.8 ± 0.3	98.9 ± 0.2	-	-	-	-
Meta Curvature	Y	99.97±0.06	99.89±0.06	99.12±0.16	99.65±0.05	-	-	-	-
DCN	Y	99.800 ± 0.050	99.891 ± 0.008	98.825 ± 0.025	99.505 ± 0.004	-	-	-	-
Prototypical Net	N	98.80	99.70	96.00	98.90	-	-	-	-
Siamse Net	N	97.30	98.40	88.20	97.00	-	-	-	-
Neural Static	N	98.10	99.50	93.20	98.10	-	-	-	-
Matching Net	N	98.10	98.90	93.80	98.50	-	-	-	-
CsTL (ours)	N	98.26±0.18	99.86±0.05	94.45±0.16	98.48±.09	90.76±0.13	97.35±.07	86.96±.11	95.50±.07

Evaluation of CsTL with different N and k values (Fig. 6.5) suggested that a bigger number of support-set produced a higher accuracy rate; the significant difference ($p < 0.05$) was observed when using 1-shot and 5-shot.

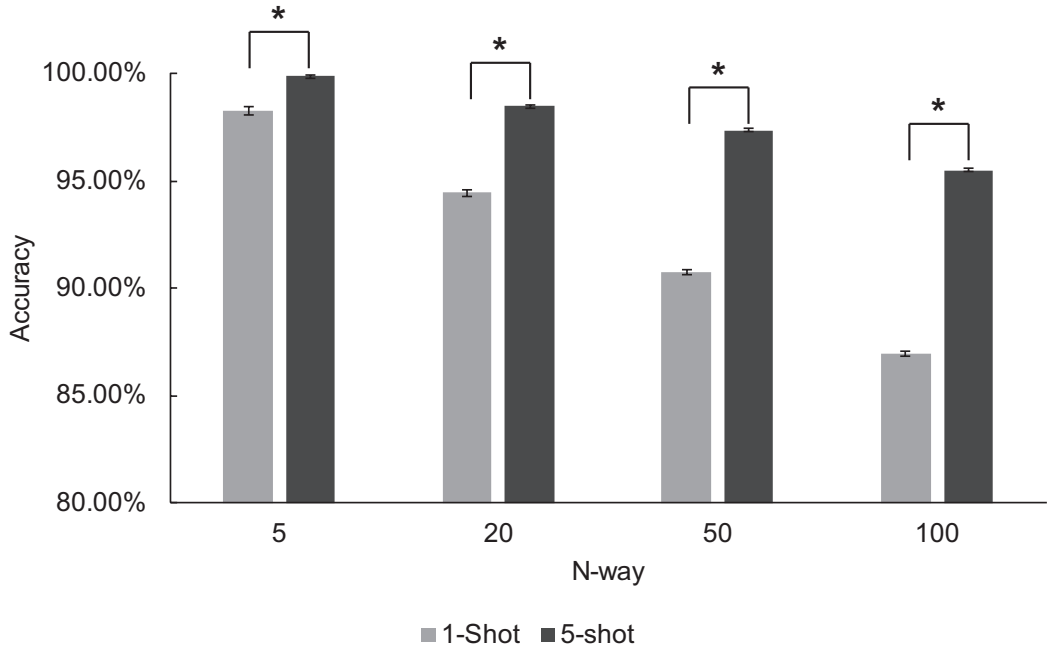


Fig. 6.5: Average accuracy rate of CsTL over 1000 episodes on Omniglot dataset using different values of N-way. A significant difference between 1-shot and 5-shot groups was computed using t-test. * indicates $p < 0.05$.

6.4.3 AttentionTest Results: Comparison

All methods attained the same accuracy rate on differentiating typical from ASD groups (Tab. 6.2). But, CsTL achieved a higher the Matthews correlation coefficient (MCC) [26] score that showed a strong positive relationship between its prediction and the ground truth labels. In recognizing new-class (adult), CsTL attained perfect accuracy rate while k-NN got 33.33 % lower accuracy rate. Besides, the accuracy rate of “All classes” demonstrated the degradation of k-NN performance when adding the new class to its support sets, which was

suggested by lower accuracy and MCC score. The same effect was not observed in CsTL's performance.

Table 6.2: Accuracy rate (%) and MCC score of CsTL and baseline on AttentionTest dataset .

Algorithm	Typical/ASD		Novel-class		Combined	
	ACC	MCC	ACC	MCC	ACC	MCC
k-NN	80.77	.61	66.67	-	79.31	.64
DNN	80.77	.60	x	x	x	x
xGBoost	80.77	.61	x	x	x	x
CsTL	80.77	.64	100.00	-	82.76	.73

6.4.4 Distribution and Decision Boundary of the Proposed System

Fig. 6.6 shows distribution of latent variables of the proposed DDL. The distribution showed a high variance of ASD children, causing a high overlap in the distribution of the second principal component.

Decision boundary of the proposed model was decided based on the local geometry of data distribution. Though the decision boundary was not smooth, the higher value of neighbour k produces smoother decision boundary (Fig. 6.7).

Considering the results of data distribution and decision boundary, the proposed diagnostic support system measured similarity score instead of class probability to identify developmental disorder in children. The similarity score represented how close an individual with the typical and ASD groups that was based on the local geometry of data distribution.

6.4.5 AttentionTest Results: Retrieval

Retrieval results of DDL demonstrated that correctly classified query yielded the first and second rank retrieved support set with the same label as the query

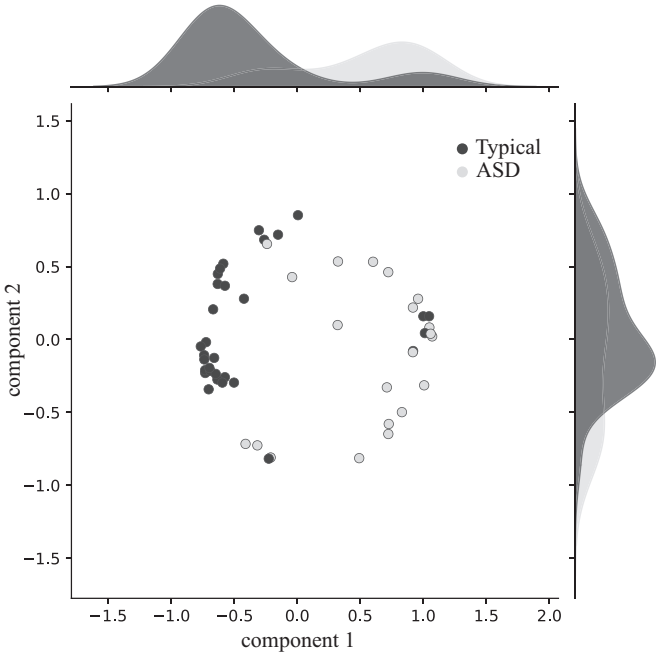


Fig. 6.6: Distribution of latent variables of typical and ASD groups. For the sake of visualization, the dimension of latent variables was reduced to two using Principal Component Analysis (PCA, explained variance ratio: 0.71).

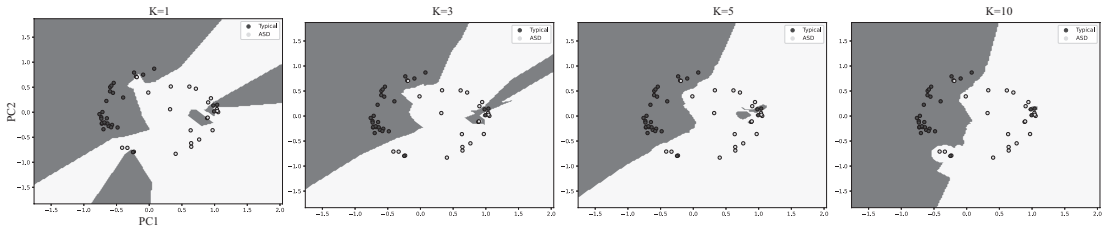


Fig. 6.7: Decision boundary of the proposed model with different values of k.

(Fig. 6.8 and 6.9). While, the retrieval results of the misclassified query had a different label from it (Fig. 6.10). The star plots also (Fig. 6.10) reveal lower mean values and variability in the misclassified ASD subjects' features than the correctly classified subject (Fig. 6.9).

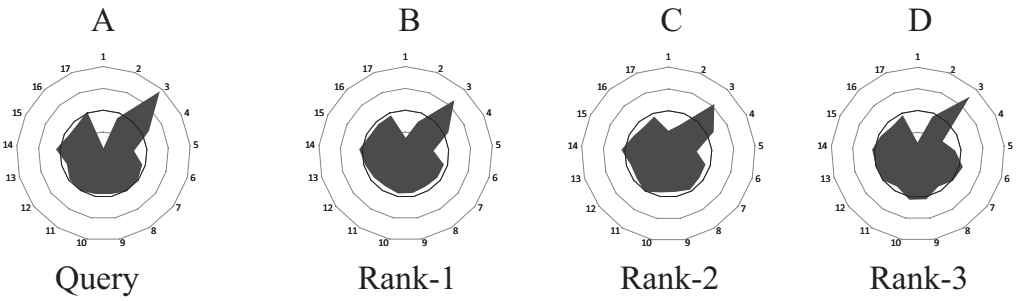


Fig. 6.8: Star plots of features. The query (A) had typical label. The 1st (B), 2nd (C) and 3rd (D) rank retrieval results were typical children.

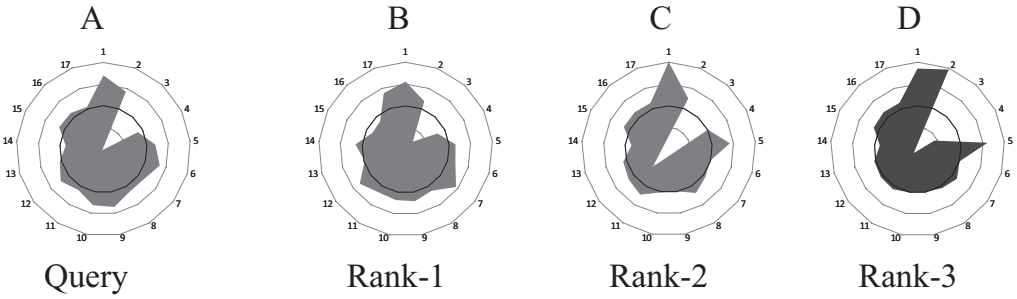


Fig. 6.9: Star plots of features. The query (A) had ASD label. The 1st (B), 2nd (C) rank retrieval results were children with ASD symptoms, while the 3rd one (D) was typical children.

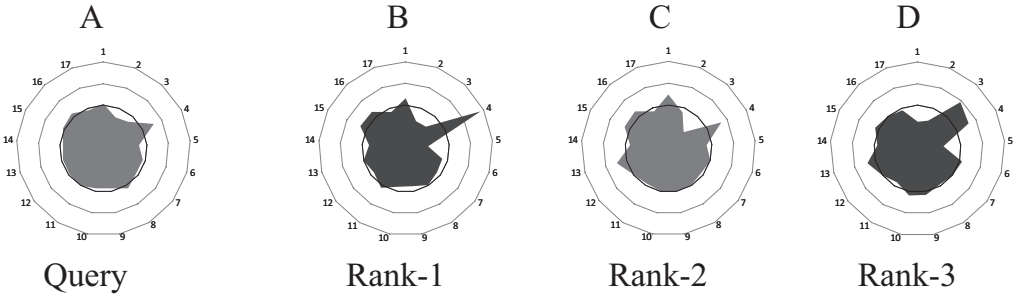


Fig. 6.10: Star plots of features. The query (A) had ASD label. The 2nd rank retrieval result was children with ASD symptoms, while the 1st (C) and 3rd (D) ones were typical children.

6.4.6 AttentionTest Results: Similarity Score and Interpretation

Similarity score measured how similar an individual was to typical and ASD groups. SHAP value explained the similarity score for each query by estimating

contribution of each feature to the score. The results signified features contribution in each query differed and they could provide comprehesive information to the users (Fig. 6.8, 6.9, and 6.10). Interpretation results of correctly classified typical child suggested that gaze-adjustment and response related features were more informative than gaze-related features. With ASD query, however, response related features (go-positive) were less informative than gaze-related features. These results indicated psychiatrists may use this interpretation to focus on examining specific symptoms as game performance and gaze modulation related to different disorder symptoms; RT-Var correlated to impulsivity and irregularity in gaze modulation might correspond to inattentiveness.

6.4.7 Preliminary Results Using Features from Group-level and Individual-level Systems

In this study, we conducted the preliminary results employing features from group-level (Chapter 4) and individual-level (Chapter 5) monitoring systems. Three children (3, 4, and 5 years) took part in the experiment. We added Gaussian noise to the features of typical children and labeled the data as non-typical.

The results of first query (Fig. 6.14 A and B) showed that features from individual-level monitoring suggested that the query belong to the typical group. The second query's results (Fig. 6.14 C and D) showed that sample entropy of moving velocity during free-playing and gaze-related features proposed it belonged to non-typical group.

6.5 Concluding Remarks

The experimental results on Omniglot datasets suggested CsTL has a promising recognition rate compared to the state-of-the-art. Its application in identifying typical, ASD, and adults demonstrated reliable performance showed by high accuracy and MCC score.

Using DDL trained with CsTL, the study could retrieve support set that

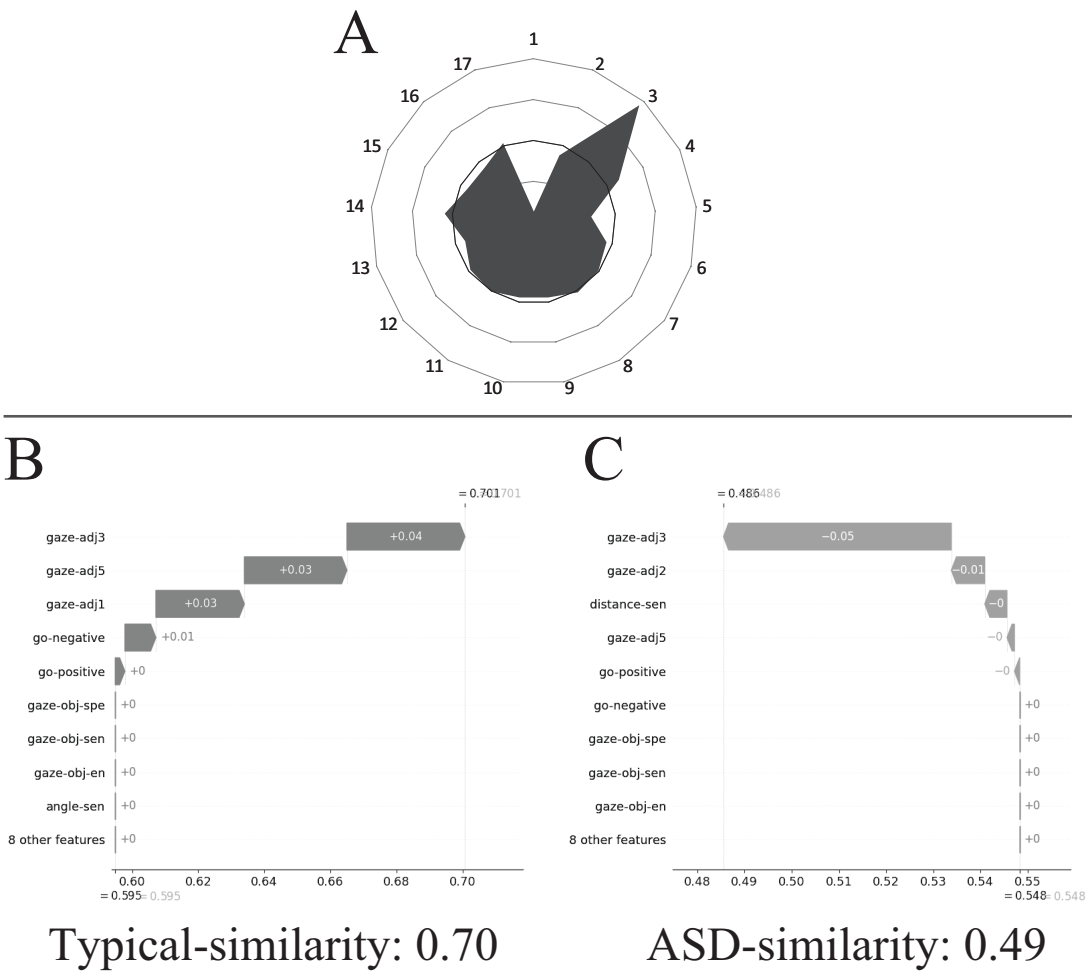


Fig. 6.11: SHAP values of similarity score for a query with typical label (A). (B) SHAP values between the query and cluster center of typical subjects. (C) SHAP values between the query and cluster center of children with ASD symptoms.

was like the query. Employing the SHAP value [94], this study interpreted the model's similarity estimation. Two limitations of this work are few benchmark studies and the small sample of the AttentionTest dataset. The proposed loss was evaluated on two benchmark datasets that are not enough to make a good assessment.

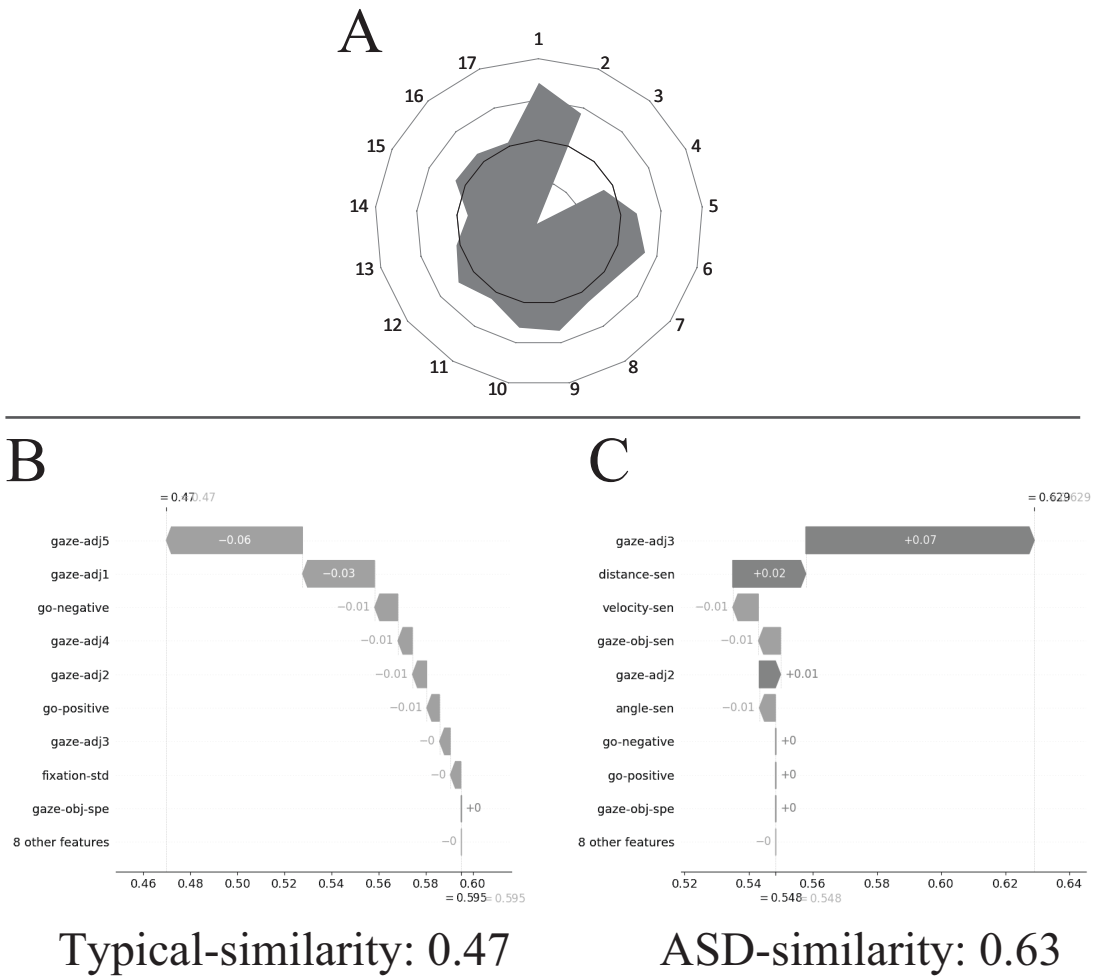


Fig. 6.12: SHAP values of similarity score for a query with ASD label (A). (B) SHAP values between the query and cluster center of typical subjects. (C) SHAP values between the query and cluster center of children with ASD symptoms.

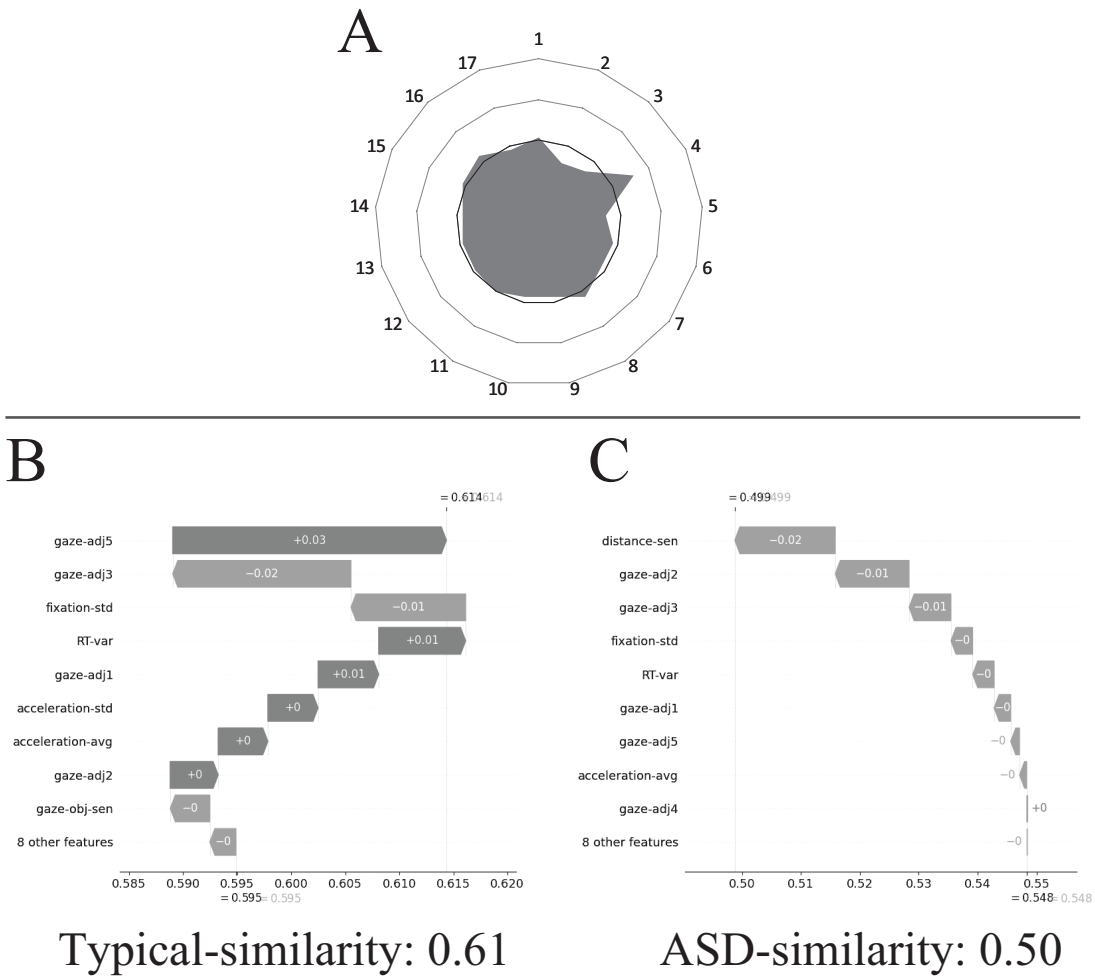
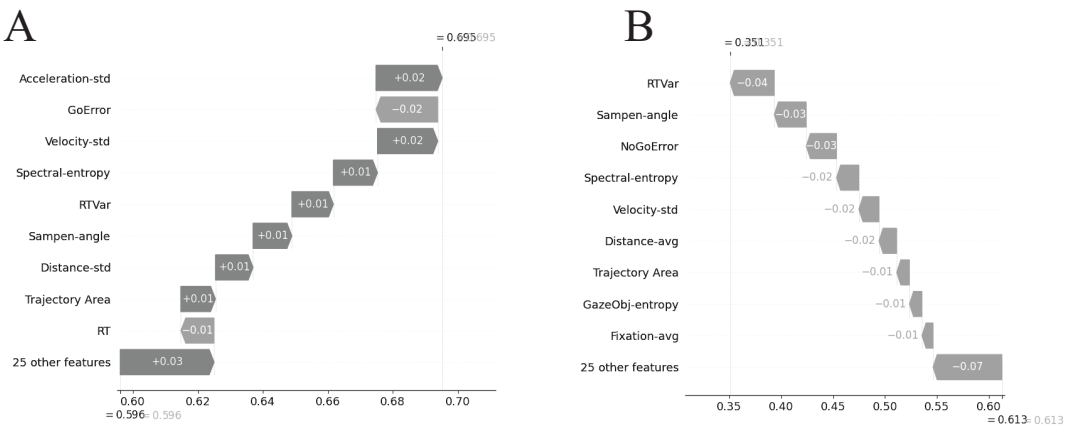
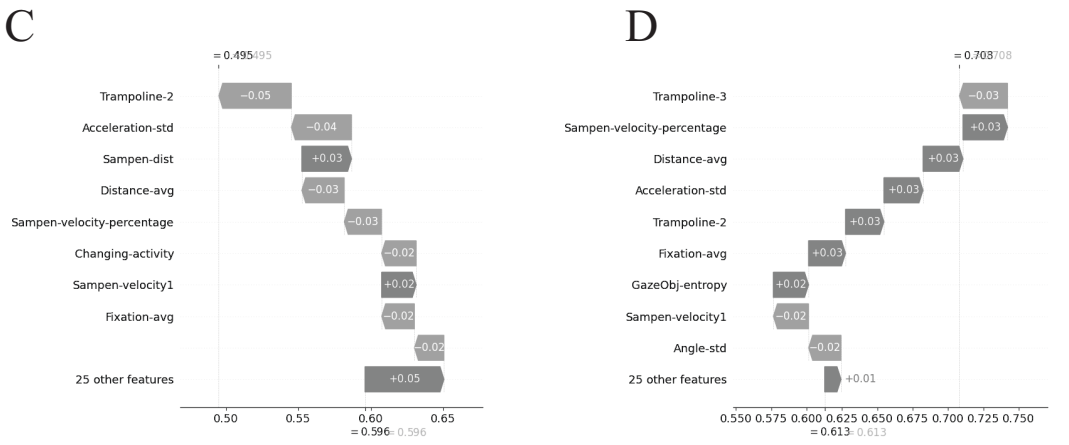


Fig. 6.13: SHAP values of similarity score for a query. (B) SHAP values between the query and cluster center of typical subjects. (C) SHAP values between the query and cluster center of children with ASD symptoms.



Typical-similarity: 0.69 Non-typical-similarity: 0.35



Typical-similarity: 0.49 Non-typical-similarity: 0.70

Fig. 6.14: SHAP values of similarity score for typical (A-B) and non-typical queries (C-D).

Chapter 7

Conclusion

This paper discusses a novel diagnostic support system to identify developmental disorder symptoms in children. The proposed system comprised group-level and individual level monitoring system. Group-level system monitored children's behavior in the nursery room by tracking children's activity with multiple Kinect sensors and RGB cameras. Individual-level system measured children's response and gaze behavior when they played a game version of the Go/NoGo task.

Chapter 1 describes the detail of the proposed system. Our decision support system employed deep distance learning (DDL) and SHAP values to provide interpretable evidence-based results. To measure similarity between a query and support sets, DDL used the combined features from group-level and individual-level monitoring system. Then, SHAP value interpreted the similarity score produced by DDL.

In Chapter 3, we address the study of Deep Neural Network (DNN) model to estimate human activity from multiple views. The proposed DNN comprised pre-trained CNNs, attention, RNN, and Softmax layers. It extracted features from multiple-view with shared-weight pre-trained VGG-16 and filtered out uninformative features using attention-layer. Afterwards, the proposed model processed the temporal information using RNN before computing the class probability with Softmax layer. Experimental results on IXMAS and i3DPost showed that the proposed model outperformed performance of conventional CV and DNN based

method using 2D inputs, and achieved competitive results compared to methods using 3D-representation and multimodal inputs [1]. The results also implied that using multiple-view input could resolve occlusion issue that often occurred in single-view application; multi-view application resulted in higher accuracy rate. Besides, online classification results suggested that longer sequence of clip yielded higher recognition rate.

Chapter 4 outlines the results of our study in quantifying children's playing behavior with marker-less method. The study used multiple Kinect sensors and RGB cameras to track children's activity in the nursery school. We utilized OpenPTrack library to perform persons and objects tracking with multiple Kinect sensors. After modeling children's behavior with PetriNet, we extracted four features from it to represent children's behavior. Statistical comparison between typical and ASD groups demonstrated that ASD children changed their activity and played alone more frequently than their typical peers did. The results, however, showed that the difference between the groups for other features was insignificant [2].

While previous chapters explain the group-level monitoring, Chapter 5 presents the investigation results of individual-level system. In this study, we developed a serious game version of the Go/NoGo task (AttentionTest) and investigate the relations between participants' response and their gaze behavior during the task. We represented respectively Go and NoGo stimulus as "chicken" and "cat" characters. Participants had to respond to the chicken character when it appeared. But, they should inhibit their action towards cat character. The proposed system tracked participants' gaze movement with Tobii eye-tracker mounted on the monitor. From subjects' response and gaze behavior, we extracted two types of features: game performance and gaze behavior. Statistical analysis results showed that significant positive relationship existed between subjects' response and their gaze behavior: participants with bigger gaze trajectory area responded slower with higher variance [3]. Clustering of those features with K-means suggested that there were two clusters in the population. The participants belong to the

first cluster, focused their gaze on the center of the screen. While the members of the second cluster adjusted their gaze to stimuli position. Also, the second cluster's member responded to the stimuli faster with higher variance than the first cluster's.

Referring to the results of our previous experiment, we conducted a study to identify ASD symptoms in children using features extracted from their response and gaze behavior when they played AttentionTest game. We recruited 35 typical and 22 children with ASD in this study. During the experiment, all participants took one minute training before participating in a four-minute evaluation. We extracted spatial and gaze-adjustment features and performed statistical comparison with Student t, Mann-Whitney U, and ANOVA test. This study conducted two statistical comparisons: between typical and ASD groups and among typical ASD with ADHD, and ASD without ADHD groups. Statistical comparison results signified that ASD children significantly differed from the typical ones in their gaze related features but not in their response towards the stimulus. The results showed higher irregularity in gaze modulation of children with ASD symptoms when they adjusted their gaze to stimulus position [4]. The results might correlate with their low ability to filter out uninformative information, causing more unintentional viewing behavior. In contrast, comparison between ASD with and without ADHD groups showed an insignificant difference. The results, however, demonstrated higher irregularity of gaze modulation among ASD subjects with ADHD comorbid.

Employing informative features that we found in the study explained in Chapter 5, we developed a decision support system using Deep Distance Learning and SHAP value algorithm. DDL allowed the proposed system to compute similarity between a query and support sets, retrieve support sets that are like the query, and perform classification of novel class; the proposed system interpreted similarity scores with SHAP value. In this study, we proposed Cluster Hard Triplet Loss (CsTL) to train the deep distance model. The loss minimized the distance between an anchor to positive cluster and maximize the distance of it to the

closest negative cluster. Evaluation with Omniglot dataset demonstrated CsTL got competitive results compared to the state-of-the-art methods. Also, the DDL outperformed k-NN and Deep Neural Network model in identifying ASD symptoms in children. Retrieval results provided evidence-based results: the first and second rank results of correctly classified query had the same label as it. SHAP value of similarity score revealed that the most informative features in differentiating typical from ASD groups were gaze-adjustment features. Besides, different in each case, game performance (RT-var, Go positive and negative scores) and gaze behavior feature might follow afterward.

Results of this study suggested that we might use marker-less method in developing a decision support system to identify developmental disorder symptoms such as ADHD and ASD in children. First, the results demonstrated that our proposed system employing OpenPTrack with multiple Kinect sensors could identify ASD/ADHD symptom-related features, such as frequency of changing activity and tendency to play alone. While the individual-level system equipped with an eye-tracker could differentiate typical of their peers with ASD/ADHD system by extracting response and gaze related features. Last, the proposed decision support system using Deep Distance Learning trained with CsTL and SHAP value could provide interpretable evidence-based results that might help psychiatrist in making better judgment to identify developmental disorder symptoms.

There are three limitations to this study. First, we have not utilized both features from group-level and individual-level monitoring system in our decision support system to identify symptoms of ASD/ADHD in children. Also, this study has not investigated the relation between features extracted using the proposed system and behavioral assessment checklist, such as the Children Behavior Checklist (CBCL). Last, it would be of interest to consider subjects across a broader age range to enable capturing a greater variety of behaviors. Future studies should jointly measure children's behavior with group and individual monitoring systems and combine features extracted from those features with behavioral checklist scores.

Bibliography

[1] P. U. Putra, K. Shima, and K. Shimatani, "Markerless human activity recognition method based on deep neural network model using multiple cameras," in *2018 5th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2018, pp. 13–18.

[2] P. Putra, K. Shima, and K. Shimatani, "Markerless behavior monitoring system for diagnosis support of developmental disorder symptoms in children," in *2021 21st International Conference on Control, Automation and Systems*. IEEE, 2021, pp. 1784–1787.

[3] P. Putra, K. Shima, and K. Shimatani, "Catchicken: A serious game based on the go/nogo task to estimate inattentiveness and impulsivity symptoms," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2020, pp. 152–157.

[4] P. U. Putra, K. Shima, S. A. Alvarez, and K. Shimatani, "Identifying autism spectrum disorder symptoms using response and gaze behavior during the go/nogo game catchicken," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[5] F. Chiarotti and A. Venerosi, "Epidemiology of autism spectrum disorders: A review of worldwide prevalence estimates since 2014," *Brain Sci.*, vol. 10, no. 5, pp. 274–294, May 2020.

[6] M. J. Maenner, K. A. Shaw, J. Baio *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2016," *MMWR Surveillance Summaries*, vol. 69, no. 4, p. 1, 2020.

[7] M. L. Danielson, R. H. Bitsko, R. M. Ghandour, J. R. Holbrook, M. D. Kogan, and S. J. Blumberg, "Prevalence of parent-reported adhd diagnosis and associated treatment among us children and adolescents, 2016," *Journal of Clinical Child & Adolescent Psychology*, vol. 47, no. 2, pp. 199–212, 2018.

[8] M. R. Mohammadi, A. Khaleghi, A. M. Nasrabadi, S. Rafieivand, M. Begol, and H. Zarafshan, "Eeg classification of adhd and normal children using non-linear features and neural network," *Biomedical Engineering Letters*, vol. 6, no. 2, pp. 66–73, 2016.

[9] X. Peng, P. Lin, T. Zhang, and J. Wang, "Extreme learning machine-based classification of adhd using brain structural mri data," *PloS one*, vol. 8, no. 11, p. e79476, 2013.

[10] S. Bezdjian, L. A. Baker, D. I. Lozano, and A. Raine, "Assessing inattention and impulsivity in children during the go/nogo task," *British Journal of Developmental Psychology*, vol. 27, no. 2, pp. 365–383, 2009.

[11] J. D. Kropotov, V. A. Grin-Yatsenko, V. A. Ponomarev, L. S. Chutko, E. A. Yakovenko, and I. S. Nikishena, "Erps correlates of eeg relative beta training in adhd children," *International journal of psychophysiology*, vol. 55, no. 1, pp. 23–34, 2005.

[12] D. M. Eagle, A. Bari, and T. W. Robbins, "The neuropsychopharmacology of action inhibition: cross-species translation of the stop-signal and go/no-go tasks," *Psychopharmacology*, vol. 199, no. 3, pp. 439–456, 2008.

[13] M. R. Swanson and M. Siller, "Patterns of gaze behavior during an eye-tracking measure of joint attention in typically developing children and children with autism spectrum disorder," *Research in Autism Spectrum Disorders*, vol. 7, no. 9, pp. 1087–1096, 2013.

[14] B. Noris, J. Nadel, M. Barker, N. Hadjikhani, and A. Billard, "Investigating gaze of children with asd in naturalistic settings," *PloS one*, vol. 7, no. 9, p. e44144, 2012.

[15] L. Meinecke, N. Breitbach-Faller, C. Bartz, R. Damen, G. Rau, and C. Disselhorst-Klug, "Movement analysis in the early detection of newborns at risk for developing spasticity due to infantile cerebral palsy," *Human movement science*, vol. 25, no. 2, pp. 125–144, 2006.

[16] T. Tsuji, S. Nakashima, H. Hayashi, Z. Soh, A. Furui, T. Shibanoki, K. Shima, and K. Shimatani, "Markerless measurement and evaluation of general movements in infants," *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.

[17] R. Migita, K. Shimatani, T. Shibanoki, Y. Kurita, K. Shima, and T. Tsuji, "A marker-less monitoring system for behavior analysis of infant using video image (in japanese)," *Japan Journal of Human Growth and Development Research*, vol. 2014, no. 65, pp. 1–7, 2014.

[18] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[19] F. Heinze, K. Hesels, N. Breitbach-Faller, T. Schmitz-Rode, and C. Disselhorst-Klug, "Movement analysis by accelerometry of newborns and infants for the early detection of movement disorders due to infantile cerebral palsy," *Medical & biological engineering & computing*, vol. 48, no. 8, pp. 765–772, 2010.

[20] M. J. Kofler, M. D. Rapport, D. E. Sarver, J. S. Raiker, S. A. Orban, L. M. Friedman, and E. G. Kolomeyer, "Reaction time variability in adhd: a meta-analytic review of 319 studies," *Clinical psychology review*, vol. 33, no. 6, pp. 795–811, 2013.

[21] G. S. Young, N. Merin, S. J. Rogers, and S. Ozonoff, "Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism," *Developmental science*, vol. 12, no. 5, pp. 798–814, 2009.

[22] S. Faja, S. J. Webb, E. Jones, K. Merkle, D. Kamara, J. Bavaro, E. Aylward, and G. Dawson, "The effects of face expertise training on the behavioral performance and brain activity of adults with high functioning autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 42, no. 2, pp. 278–293, 2012.

[23] R. Beaumont and K. Sofronoff, "A multi-component social skills intervention for children with asperger syndrome: The junior detective training program," *Journal of Child Psychology and Psychiatry*, vol. 49, no. 7, pp. 743–753, 2008.

[24] P. J. Prins, E. T. Brink, S. Dovis, A. Ponsioen, H. M. Geurts, M. De Vries, and S. Van Der Oord, "" braingame brian ": toward an executive function training program with game elements for children with adhd and cognitive control problems," *GAMES FOR HEALTH: Research, Development, and Clinical Applications*, vol. 2, no. 1, pp. 44–49, 2013.

[25] P. Vonbach, A. Dubied, S. Krähenbühl, and J. H. Beer, "Prevalence of drug–drug interactions at hospital entry and during hospital stay of patients in internal medicine," *European journal of internal medicine*, vol. 19, no. 6, pp. 413–420, 2008.

[26] J. A. Lipton, R. J. Barendse, A. F. Schinkel, K. M. Akkerhuis, M. L. Simoons, and E. J. Sijbrands, "Impact of an alerting clinical decision support system for glucose control on protocol compliance and glycemic control in the intensive cardiac care unit," *Diabetes technology & therapeutics*, vol. 13, no. 3, pp. 343–349, 2011.

[27] S. T. McMullin, T. P. Lonergan, C. S. Rynearson, T. D. Doerr, P. A. Veregge, and E. S. Scanlan, "Impact of an evidence-based computerized decision support system on primary care prescription costs," *The Annals of Family Medicine*, vol. 2, no. 5, pp. 494–498, 2004.

[28] E. S. Berner, *Clinical decision support systems*. Springer, 2007, vol. 233.

[29] L. Oakden-Rayner, G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, and L. J. Palmer, "Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[30] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

[31] A.-A. Liu, N. Xu, Y.-T. Su, H. Lin, T. Hao, and Z.-X. Yang, "Single/multi-view human action recognition via regularized multi-task learning," *Neurocomputing*, vol. 151, pp. 544–553, 2015.

[32] C. Torres, V. Fragoso, S. D. Hammond, J. C. Fried, and B. Manjunath, "Eye-cu: Sleep pose classification for healthcare using multimodal multiview data," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9.

[33] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *European Conference on Computer Vision*. Springer, 2010, pp. 635–648.

[34] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[35] Z. Gao, H.-Z. Xuan, H. Zhang, S. Wan, and K.-K. R. Choo, "Adaptive fusion and category-level dictionary learning model for multiview human action recognition," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9280–9293, 2019.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[37] R. Kavi, V. Kulathumani, F. Rohit, and V. Kecojevic, "Multiview fusion for activity recognition using deep neural networks," *Journal of Electronic Imaging*, vol. 25, no. 4, pp. 043 010–043 010, 2016.

[38] D. Wang, W. Ouyang, W. Li, and D. Xu, "Dividing and aggregating network for multi-view action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 451–467.

[39] M. A. Khan, M. Sharif, T. Akram, M. Raza, T. Saba, and A. Rehman, "Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition," *Applied Soft Computing*, vol. 87, p. 105986, 2020.

[40] M. Gnouma, A. Ladjailia, R. Ejbali, and M. Zaied, "Stacked sparse autoencoder and history of binary motion image for human activity recognition," *Multimedia Tools and Applications*, vol. 78, no. 2, pp. 2157–2179, 2019.

[41] D. Purwanto, R. Renanda Adhi Pramono, Y.-T. Chen, and W.-H. Fang, "Extreme low resolution action recognition with spatial-temporal multi-head self-attention and knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[42] Y. Gu, X. Ye, W. Sheng, Y. Ou, and Y. Li, "Multiple stream deep learning model for human action recognition," *Image and Vision Computing*, vol. 93, p. 103818, 2020.

[43] H. Hwang, C. Jang, G. Park, J. Cho, and I.-J. Kim, "Eldersim: A synthetic data generation platform for human action recognition in eldercare applications," *arXiv preprint arXiv:2010.14742*, 2020.

[44] J. Zheng, Z. Jiang, and R. Chellappa, "Cross-view action recognition via transferable dictionary learning," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2542–2556, 2016.

[45] S. Vyas, Y. S. Rawat, and M. Shah, "Multi-view action recognition using cross-view video prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 2020, pp. 427–444.

[46] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[47] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," in *Visual Media Production, 2009. CVMP'09. Conference for*. IEEE, 2009, pp. 159–168.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[49] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[51] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[52] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 1990.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] X. Li, Z. Zhou, L. Chen, and L. Gao, "Residual attention-based lstm for video captioning," *World Wide Web*, vol. 22, no. 2, pp. 621–636, 2019.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[57] J. Kim, M. El-Khamy, and J. Lee, "Residual lstm: Design of a deep recurrent architecture for distant speech recognition," *arXiv preprint arXiv:1701.03360*, 2017.

[58] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6a overview of mini–batch gradient descent," *Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online*, 2012.

[59] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[61] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[62] S. Pehlivan and P. Duygulu, "A new pose-based representation for recognizing actions from multiple cameras," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 140–151, 2011.

[63] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on stiefel and grassmann manifolds with applications in computer vision," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8.

[64] M. B. Holte, B. Chakraborty, J. Gonzalez, and T. B. Moeslund, "A local 3-d motion descriptor for multi-view human action recognition from 4-d spatio-temporal interest points," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 553–565, 2012.

[65] S. Spurlock and R. Souvenir, "Dynamic view selection for multi-camera action recognition," *Machine Vision and Applications*, vol. 27, no. 1, pp. 53–63, 2016.

[66] K. K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," in *Computer vision, 2009 ieee 12th international conference on.* IEEE, 2009, pp. 1010–1017.

[67] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *CVPR 2011.* IEEE, 2011, pp. 3209–3216.

[68] N. Käse, M. Babaee, and G. Rigoll, "Multi-view human activity recognition using motion frequency," in *Image Processing (ICIP), 2017 IEEE International Conference on.* IEEE, 2017, pp. 3963–3967.

[69] D. T. Tran, H. Yamazoe, and J.-H. Lee, "Multi-scale affined-hof and dimension selection for view-unconstrained action recognition," *Applied Intelligence*, pp. 1–19, 2020.

[70] V. Mygdalis, A. Tefas, and I. Pitas, "Exploiting multiplex data relationships in support vector machines," *Pattern Recognition*, vol. 85, pp. 70–77, 2019.

[71] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2d pose-based real-time human action recognition with occlusion-handling," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1433–1446, 2019.

[72] A. P. Association *et al.*, "Diagnostic and statistical manual of mental disorders," *BMC Med*, vol. 17, pp. 133–137, 2013.

[73] M. Carraro, M. Munaro, J. Burke, and E. Menegatti, "Real-time marker-less multi-person 3d pose estimation in rgb-depth camera networks," in *International Conference on Intelligent Autonomous Systems.* Springer, 2018, pp. 534–545.

[74] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: real-time multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

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

[76] E. Best and R. Devillers, "Sequential and concurrent behaviour in petri net theory," *Theoretical Computer Science*, vol. 55, no. 1, pp. 87–136, 1987.

[77] T. M. Cover, *Elements of information theory.* John Wiley & Sons, 1999.

[78] S. L. Hyman, S. E. Levy, S. M. Myers *et al.*, "Identification, evaluation, and management of children with autism spectrum disorder," *Pediatrics*, vol. 145, no. 1, 2020.

[79] J. R. Magnuson, G. Iarocci, S. M. Doesburg, and S. Moreno, "Increased intra-subject variability of reaction times and single-trial event-related potential components in children with autism spectrum disorder," *Autism Research*, vol. 13, no. 2, pp. 221–229, 2020.

[80] R.-A. Müller, N. Kleinhans, N. Kemmotsu, K. Pierce, and E. Courchesne, "Abnormal variability and distribution of functional maps in autism: an fmri study of visuomotor learning," *American Journal of Psychiatry*, vol. 160, no. 10, pp. 1847–1862, 2003.

[81] Z. Zhao, H. Tang, X. Zhang, X. Qu, X. Hu, J. Lu *et al.*, "Classification of children with autism and typical development using eye-tracking data from face-to-face conversations: Machine learning model development and performance evaluation," *Journal of Medical Internet Research*, vol. 23, no. 8, p. e29328, 2021.

[82] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[83] Q. He, Q. Wang, Y. Wu, L. Yi, and K. Wei, "Automatic classification of children with autism spectrum disorder by using a computerized visual-orienting task," *PsyCh Journal*, 2021.

[84] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[85] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[86] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[87] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[88] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[89] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[90] G. M. Sullivan and R. Feinn, "Using effect size—or why the p value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, p. 279, 2012.

[91] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari, "Behavioral manifestations of autism in the first year of life," *International journal of developmental neuroscience*, vol. 23, no. 2-3, pp. 143–152, 2005.

[92] B. A. Shiferaw, L. A. Downey, J. Westlake, B. Stevens, S. M. Rajaratnam, D. J. Berlowitz, P. Swann, and M. E. Howard, "Stationary gaze entropy predicts lane departure events in sleep-deprived drivers," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[93] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[94] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[95] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[96] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[97] A. Li, T. Luo, T. Xiang, W. Huang, and L. Wang, "Few-shot learning with global class representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9715–9724.

[98] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[99] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[100] H. Edwards and A. Storkey, "Towards a neural statistician," *arXiv preprint arXiv:1606.02185*, 2016.

[101] J. Liu, S. J. Gibson, and M. Osadchy, "Learning to support: Exploiting structure information in support sets for one-shot learning," *arXiv preprint arXiv:1808.07270*, 2018.

[102] Ł. Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," *arXiv preprint arXiv:1703.03129*, 2017.

[103] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[104] E. Park and J. B. Oliva, "Meta-curvature," *arXiv preprint arXiv:1902.03356*, 2019.

[105] J. Liu, F. Chao, L. Yang, C.-M. Lin, and Q. Shen, "Decoder choice network for meta-learning," *arXiv preprint arXiv:1909.11446*, 2019.

Published Papers

Peer-reviewed journals

- Putra, Prasetia Utama, Keisuke Shima, Sergio A. Alvarez, and Koji Shimatani. "Identifying autism spectrum disorder symptoms using response and gaze behavior during the Go/NoGo game CatChicken." Scientific Reports 11.1: 1-12, 2021.
- Putra, Prasetia Utama, Keisuke Shima, and Koji Shimatani. "A deep neural network model for multi-view human activity recognition ." PlosOne 17.1: 1-20, 2022.

International Conference

- Putra, Prasetia Utama, Keisuke Shima, Sayaka Hotchi and Koji Shimatani. "Markerless Behavior Monitoring System for Diagnosis Support of Developmental Disorder Symptoms in Children." Proceedings of 21st International Conference on Control, Automation and Systems, pp 1784-1787, 2021.
- Putra, Prasetia Utama, Keisuke Shima and Koji Shimatani. "Catchicken: A Serious Game Based on the Go/NoGo Task to Estimate Inattentiveness and Impulsivity Symptoms." Proceedings of 33rd International Symposium on Computer-Based Medical Systems, pp 152-157, 2020.
- Putra, Prasetia Utama, Keisuke Shima and Koji Shimatani. "Markerless Human Activity Recognition Method Based on Deep Neural Network Model Using Multiple Cameras ." Proceedings of 5th International Conference on Control, Decision and Information Technologies, pp 13-18, 2018.