Identifying Children Autism Spectrum Disorder via Machine Learning based Behavior Analysis

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Wenbo Liu

B.S., Information Engineering, South China University of Technology
M.S., Communications and Information system, South China University of Technology

Carnegie Mellon University
Pittsburgh, PA

September 2022

Acknowledgements

First and foremost, I want to thank my advisors Prof. Ming Li and Prof. Bhiksha Raj who provided me the precious opportunity to pursue PhD. Without any exaggeration, it was this period of study that completely changed the course of my life. Because of this experience, I was able to receive the best education in the world, get in touch with state-of-the-art research, train my mind to think rigorously, and start my career in AI-related areas. Ming and Bhiksha have shown me, both consciously and unconsciously, what a PhD student should be. And I often feel privileged becoming their student.

I shaped my PhD direction under Ming's guidance. It was his great initiative that allowed me to delve into autism research, define research problems, and build up the data collection environment and team in China. It was a long journey with many challenges. I still remember the time when we repeatedly traveled between Shunde and Guangzhou to build up the lab and meet with different doctors for task design. Without Ming's guidance, I would not be able to achieve what I had today.

My PhD study at CMU has received constant supports from Bhiksha. He is not only an academic advisor but is also like a real father in our student life. He likes to call us kids and always provides precious supports during our difficult times. He is the smartest person I have ever met, providing me lots of good insights and ideas. As a result, my PhD research has greatly benefited from Bhiksha's rich knowledge in machine learning.

I also want to give my thank to Carnegie Mellon University fellowship, this work was partially supported by Carnegie Mellon University through a first-year fellowship.

I would like to thank all the research collaborators along the way. My thank first goes to Prof. Li Yi at Peking University. It was the initial collaboration with Prof. Li Yi that opened a new world for me in autism research. I want to thank to Doctor Xiaobing Zou at The Third Affiliated Hospital of Sun Yat-sen University. As a renowned expert of children autism in China, Doctor Xiaobing Zou is willing to share his experience and expertise to assist our work despite his busy schedule. His help is a critical backup of our project from a clinical perspective, and we would not be able to establish the current protocol, lab environment and data collection without such backup. I would also like to thank other staffs who joined the effort, including Yixiang Xie, Xiaoyan Chen and Fengfei Zhu. Without their help, our data collection

effort could not progress so well.

I want to take this chance to thank my thesis committee members Prof. Rita Singh and Prof. Shri Narayanan. Their wisdom and advice motivated me to rethink the impact of my PhD research from a higher perspective. Their patience and kind encouragements helped me to overcome all the challenges and form the thesis into a unified set of stories.

Last but not least, I want to thank my beloved families. My husband Zhiding Yu was the reason why I initially came to US. Being another PhD student at CMU, he always inspired me with his excellence. It was his love and encouragement that led to the start of my journey. I want to express my deepest gratitude to my parents and parents in law. Their unconditional support let me pursue my dream and career more confidently. Finally, my little boy Andrew is the greatest gift I received in my PhD study. His birth brought me a deeper understanding of the significance of my research. Being a parent myself, I know what a child means to every parent. I sincerely hope my research can ultimately help those families with children that suffer from ASD.

Overall, My PhD life is a long but interesting journey. There are both highlights and lowlights, but every experience is a cherished memory. I deeply appreciate the time and efforts spent on this study. It is such a unique journey and it made me more complete.

Abstract

Autism Spectrum Disorder (ASD) is a group of lifelong neurodevelopmental disorders with complicated causes. A key symptom of ASD patients is their impaired interpersonal communication ability. In this thesis, we consider the problem of screening/diagnosing children autism spectrum disorder by analyzing their behaviors through machine learning.

Our work is motivated by the broad spectrum of previous research which indicates that children with ASD are often characterized by certain behaviors such as abnormal visual attention, lack of response to names, and impaired interpersonal communication abilities. Such behavior-level signs motivate us to analyze and identify these abnormalities under a variety of different modalities with data-driven approaches. Specifically, we begin by identifying ASD children based on their face scanning patterns. We consider using a bag-of-words (BoW) model to encode face scanning patterns, and further propose a novel dictionary learning method via discriminative mode seeking to improve the BoW representation and the identification accuracy.

To render more natural and spontaneous children reactions, we further consider an interactive diagnostic procedure under a multi-camera, multimodal system where children activities are recorded with minimum constraints. Three assessment protocols originated from the Autism Diagnostic Observation Schedule-Generic (ADOS-G) are designed: 1. response to name, 2. separation and reunion, 3. response to non-social sound stimulation. We comprehensively analyze the children behaviors through a number of computer vision, speech processing, and general machine learning approaches. Some typical problems we consider include preprocessing steps such as person detection/re-identification, pose estimation, as well as feature extraction/score prediction on top of preprocessing.

Comprehensive experimental results show that the proposed frameworks not only can effectively identify ASD, but also help human diagnosis by providing an auxiliary view with mid-level machine features/scores. Although the proposed work is yet too preliminary to directly replace existing autism assessment methods in clinical practice, it shed light on future applications of machine learning methods in early screening of the disease.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> *"To understand and measure emotional qualities is very difficult. Psychologists and educators have been struggling with that problem for years but we are still unable to measure emotional and personality traits with the exactness with which we can measure intelligence."*
>
> — Rose Zeligs, 1942

Autism Spectrum Disorder (ASD) is a pervasive developmental disorder affecting as many as 1 in 44 in the United States[1]. Children with ASD often suffer from significant impairment in social[2], occupational and other important areas of functioning. A child's autism thus can significantly affect every member of the family in terms of the additional stress on responsibilities, marriage, personal relationships, work, finance, etc. So far, there is no existing drugs cures to autism, while early intervention is known to be the best remediation for ASD children and the promptness of ASD diagnosis plays a key role in maximizing the intervention gain. Despite the fact that existing assessment methods show high validity, current ASD diagnostic approaches are both time and labour consuming. In particular, the diagnostic instruments have been designed to measure impairments mainly in: (1) language and communication; (2) reciprocal social interactions; and (3) restricted, repetitive behaviors[3]. The most widely used instruments for screening include the Autism Diagnostic Observation Schedule-Generic (ADOS-G) and the revised version ADOS-2 [4]. The subjective nature of these approaches require the accompany and administration of clinically trained professionals and the whole process can take up to 90 minutes. Such examination strategy requires

significant amount of labour and high costs, which increases the economic burdens of ASD families and lowers their chances of early diagnosis.

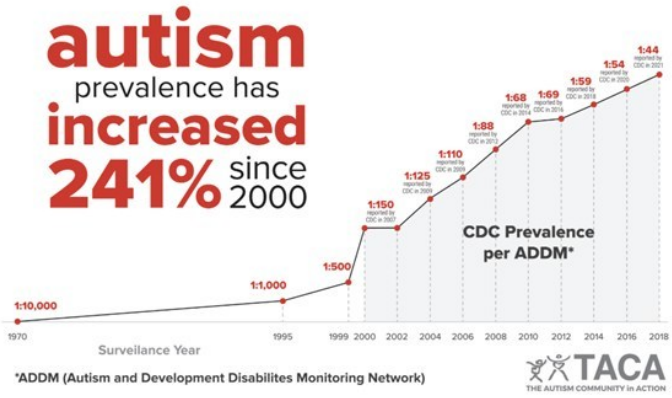

Figure 1.1: The increasing ASD rates in the past decade.

1.1 Research Theme and Motivation

Since 80 years ago, psychologists and educators had been working on understanding and measuring human's emotion and personality traits for years. How to understand human's behavior with the exactness is a challenge problem [5]. Upon witnessing the difficult lives and heavy burdens faced by the family of ASD children, we also seek to understand the child's behavior and early screen the ASD. With the recent fast development in machine learning, computer vision and speech processing, other communities such as psychology and health care benefit considerably from these booming technologies and methods. We believe that the diagnostic procedure for autism children can be decomposed into a set of activity items where one is able to quantitatively analyze certain behaviors of the children by machine learning methods to generate automatic diagnostic predictions. The purpose of this work aims at addressing the ASD diagnosis problem through multimodal behavior analysis, where the speech communications, ego-centric visual patterns, interactive activities, biosignals, and emotion states of a subject are comprehensively analyzed with state-of-the-art techniques. With the help of all these techniques, we can generate automatic diagnostic predictions and reduce intermediate human-in-loop steps.

Although reliable autism diagnosis poses a difficult problem for machine learning based techniques under any single modality, we expect that the ensemble of multiple modalities embed much richer infor-

mation to benefit our diagnosis task. We also hope that the large scale of data, together with deep learning methods, will further increase the diagnosis reliability. Our ultimate goal of this project is to provide a children-friendly ASD diagnosis environment, as well as to establish a reliable intelligent system that can objectively predict the potential ASD risk at an early stage, such that the gain of early intervention is maximized. In addition, we also seek to establish a large scale database with comprehensive modalities to benefit future autism research, children emotion analysis, and psychology studies.

1.2 Considered Research Problems

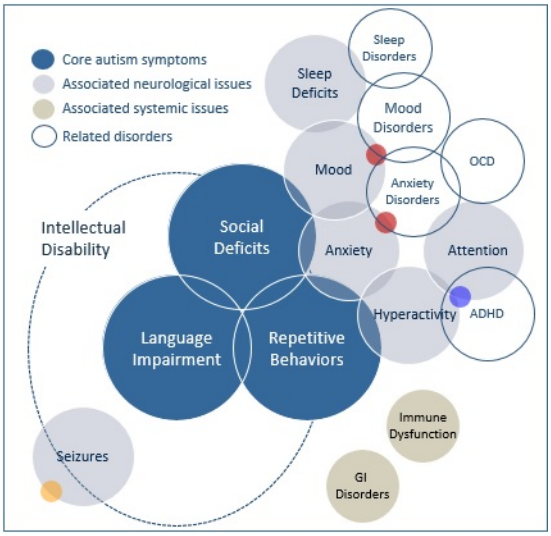

Figure 1.2: The basic symptoms of ASD.

Autism spectrum disorders are characterized by social-interaction difficulties, communication challenges and a tendency to engage in repetitive behaviors. As a result, starting from the phenomenon, we will designed the experiments on the following 4 behaviors.

The first one is visual attention analysis. Behavioral studies found that children with ASD often lack effective communications and show reduced visual attentions to faces compared to their typically developed (TD) counterparts. Existing studies further indicate that they have impairments in recognizing human faces, leading to atypical scanning patterns when looking at face images.

Secondly, we also observed that ASD children are characterized by their impaired interpersonal communication abilities. As a result, we proposed a ASD Diagnosis under multimodal, multi-Camera System.

In this system, we proposed a multimodal data collection environment, which would give children enough spaces to interact naturally with multiple participants and have more natural interactions. These natural interactions are expected to lead to more spontaneous behaviors and more accurate diagnosis. We will detailedly introduce the 2 experiments and discuss the future work of ASD early screening with multi-modality methods.

1.3 Related Work

Our work is related to or partly inspired by a wide variety of previous work, ranging from psychology, psychiatry, behavior analysis to machine learning. Below we give a brief summary of these work.

1.3.1 Psychology Study and Clinical Diagnosis for Autism

The term "autism" was first used by Paul Eugen Bleuler to define the symptoms of schizophrenia in 1912 [6]. The term "autistic psychopathy" was further followed by Hans Asperger to describe child psychology in 1938, which is now known as the Asperger syndrome [7]. In 1943, Leo Kanner introduced "early infantile autism" to describe a type of schizophrenia that causes children to lose the ability to "form the usual, biologically provided affective contact with people" [2]. Both Hans Asperger and Leo Kanner are considered to have formed the basis of the modern study of autism.

ASD often shows signs in the first 3 years of life. These signs include difficulties with both verbal and non-verbal interactions, such as the appropriate use of spoken language, eye contact, facial expression, and body gestures [8]. Other typical symptoms include social challenges in expressing or understanding emotions/intentions, as well as restricted and receptive behaviors. These core symptoms form the bases for ASD diagnosis, leading to a variety of ASD screening and diagnosis protocols. Diagnosing ASD is challenging, often requiring time-consuming and careful observation and evaluation [9]. In this process, comprehensive clinical assessments are conducted by licensed professionals on various domains, such as behavior excesses, communication, self-care, and social skills. As of today, the Diagnostic and Statistical Manual fourth edition (DSM-IV) is considered the most widely used criteria for ASD diagnosis [8]. Other popular clinical and self-screening methods include Autism Diagnostic Interview-Revised (ADI-R),

Autism Diagnostic Observation Schedule (ADOS), Childhood Autism Rating Scale (CARS), Joseph Picture self-concept scale, and the social responsiveness scale [10, 11, 12, 13]. Although proven to be effective, these methods are considered time consuming and require licensed clinicians and observers to administer the process [14, 15, 16].

1.3.2 Behavior Analysis and Machine Learning for Autism Research

The observation of abnormal visual attention in ASD children has been supported by works in psychology, psychiatry and behavior analysis. It is known that individuals with ASD have difficulties recognizing faces and interpreting facial emotions [17, 18, 19, 20, 21, 22]. When presented with human faces, they tend to show reduced visual attention [23]. Some studies have further found that ASD individuals show reduced attention to core features of faces such as eyes, nose and mouth [24, 25, 26, 27, 28, 29]. To better study the source of such difference, Yi et al. [30] introduced two control groups of age-matched and IQ-matched typically developing (TD) children. The study shows that the scanning patterns of ASD children differ from those of both TD groups significantly in the eye region. Another two studies [31, 32] reveal that individuals with ASD exhibit influences from the different race of faces. Besides human faces, it is also found that ASD individuals show atypical visual attention to general objects and natural scenes [33, 34].

An important technique to analyze the face scanning patterns in the above literature is the area of interest (AOI) [35, 36] approach. Specifically, subjects are shown with human face images on the screen and their eye movement patterns are captured by eye tracking devices. The viewed images are manually partitioned into semantically meaningful regions (eye, nose and mouth, etc.), with the frequency (counts) of eye fixations falling into each region counted and analyzed. Besides AOI, another method towards analyzing visual attentions is the iMap approach [37], where a heat map of visual attention is generated by smoothed eye fixations. This heat map indicate the density of eye-gaze coordinates, and is able to capture more fine-grained information than AOI [30]. A variety of studies have reported the application of AOI and iMap in analyzing abnormal attention in ASD [38, 39, 31, 32].

The discoveries in ASD behavior analysis have also motivated researchers to also use machine learning methods. The procedure of ASD diagnosis involves the wide understanding of human states such as emo-

tion, attention, action and reaction. It is therefore highly related to areas such as affective computing [40], speech recognition [41], action recognition [42], multimodal interaction [43], as well as general computer vision and machine learning techniques. Numerous studies have applied machine learning to autism research in behavior observation and brain activities analysis [44, 45, 46, 47, 48, 49]. Some studies have also used machine learning to effectively select a subset of features from large amounts of existing features to reduce the time of ASD diagnosis. For instance, Kosmicki et al. [46] proposed a machine learning based feature selection framework to reduce the number of behavioral features and measurements in ADOS, Duda et al. [45] used machine learning to train a classifier which can reduce 72% length in ADOS-G test while retain 97% accuracy. In general, these methods only consider using machine learning in selecting features but not in the loop of generating them. Obtaining the features thus still requires considerable human interaction. In addition, Bone et al. [50], pointed out that feature selection method is difficult to reproduce similar results on a larger and balanced dataset.

Recent autism studies start to address prediction tasks besides feature selection [44, 48, 49]. For example, a machine learning framework was proposed to classify low function children with ASD based on their motor pattern [44]. Abnormal visual attention cues have also been used to identify ASD, where deep neural networks are used to encode discriminative features from visual attention [51] and [52]. Another related literature is the work proposed by Bidwell and Essa et al. [53]. The author established a dataset containing 50 recorded "response to name" sessions, and explored markerless child head tracking with a camera recording from the top. The proposed method marks an import effort towards machine learning based human behavior analysis. These works not only motivate us to take both the activity cues and the ego-centric visual patterns as inputs to predict the ASD risk, but also show the feasibility of ASD diagnosis with multimodal signal and the complementary of multiple different assessments.

1.3.3 General Machine Learning Research

Some data-driven approaches in behavior analysis are inherently connected to general machine learning methods. For example, the AOI approach is highly related to the well-known bag-of-words (BoW) model [54] in machine learning, where counting the frequency is essentially feature encoding with his-

togram and the region partitioning corresponds to dictionary words (or codebook) in BoW. Besides AOI, using Gaussian smoothed heatmap to approximate the density of eye fixation in iMap [37] coincides with kernel density estimation [55, 56] in machine learning.

Our proposed approaches in Chapter 2 are motivated by the above observation. At a high level, our method aims to predict ASD by encoding eye movement data into features following the BoW representations. Thus an important machine learning problem this work addresses is to learn discriminative dictionary words that gives good BoW representations. Our work is in line with the works that add discriminative label information to dictionary learning [57, 58, 59]. Another heuristic to incorporate discriminative information is to learn class-specific dictionary by performing k-means in a class-wise manner [60]. Other related methods include the descriptive word ranking [61] as well as large-margin clustering using SVM and iterative cross-validation [62, 63]. Finally, one related work which partly inspired our proposed framework is [64], where the authors discover mid-level image patches with discriminative mode seeking, and formulate the mode seeking process as a constrained optimization problem. However, these works are not directly applicable to our task without significant changes.

Our proposed data processing pipelines and behavior analysis approaches in Chapter 3 and 4 involve a number of deep learning based visual recognition techniques ranging from pose estimation and tracking [65, 66, 67], face detection and alignment [68], person re-identification and action recognition [69].

1.3.4 Research impact

Our research on ASD pre-screening using abnormal attention has motivated numerous other related works. For example, [51] also considers ASD prediction by analyzing eye gaze patterns on natural images. Similarly, [70] proposes an ASD prediction framework based on eye gaze by letting the subjects watch a 10-second video of a female speaking. In addition, [71] proposes an open ASD prediction dataset containing the eye movements of children when watching natural images. Despite the slightly different settings in experiment protocols, these works follow a common assumption from the proposed research - eye gaze pattern contains rich ASD-related information that can be effectively captured.

Besides abnormal attention discovery, our application of machine learning for ASD pre-screening has

motivated numerous works that similarly propose machine learning based ASD pre-screening. For example, [72] proposes to identify ASD with fMRI using with an autoencoder and a single layer perceptron. Another work [73] proposes identify ASD with fMRI using random forest. The authors of [74] adopt linear discriminant analysis (LDA) and k-nearest neighbor (KNN) in ASD classification. Despite the difference in the proposed specific approaches, these works also follow a common assumption that machine learning is capable to recognize the underlying ASD patterns from the input data.

Finally, there have also been some concurrent works on ASD pre-screening using multimodal behavior analysis. For example, [75] conducts machine learning analysis on 3-minute home videos of children to identify ASD. [76] proposes the automatic detection of ASD from speech using deep learning techniques. Finally, a recent work [77] tries to model the dynamic relation between speech production and facial expression in ASD children, by having subjects watching and repeating spoken sentences with accompanying facial expressions. These works show the importance of multimodal analysis in ASD pre-screening and study, which is aligned with the direction of this work.

1.4 Our contributions

The contributions of this work can be summarized as follows:

- **Problem.** Our research introduced a paradigm for machine learning based ASD pre-screening, where we show how features and other useful information can be extracted by machines to help a clinician. Our work show promising results indicating that current state-of-the-art machine learning techniques are capable to model and recognize ASD-related patterns from various types of input.

- **Dataset.** We established a multi-camera, multimodel human behavior dataset with expert labeled response scores and ASD diagnosis results. With rich audio-visual recordings of spontaneous children activities and multi-person interactions, the dataset provides a good research playground for vision and multimodal learning methods.

- **Approach.** Our research involves a novel machine learning framework with state-of-the-art prediction performance for ASD pre-screening using eye tracking. Towards ASD pre-screening based on

multimodal behavior analysis, we propose a set of machine learning based preprocessing pipelines that automate the extraction of mid-level features/results. We show that our prediction frameworks, built on top of these pipelines, generate machine scores that are consistent with human experts. Finally, we also demonstrate improved prediction performance by exploring model ensemble and further demonstrate how machine prediction can be combined to further assist human screening.

Chapter 2

ASD screening via abnormal attention discovery

2.1 Motivation and Problem Statement

Eye movements encode rich information about the attention distribution and cognitive strategies during face viewing that may indicate the potential risk of ASD, such as the fixation durations and counts at different facial areas, the speed and direction of the saccades, as well as the temporal information of the face scanning pattern. Automatically handling eye gaze data with machine learning methods makes the prediction process more scalable than manually doing so. The purpose of this chapter is to propose a machine learning framework which learns from the observed face scanning patterns to automatically identify children with ASD. We hope that such a framework can generate useful mid-level features in the ASD evaluation, and that by adopting eye movement, subjective factors can be reduced to make the ASD evaluation a more objective process.

In our work, we focus on analyzing face scanning patterns to predict ASD. We adopted the eye movement dataset from a previous published work, which asked children with and without ASD to recognize Asian and Caucasian face while their eye movements being tracked. Our major contribution in this work is that we propose a machine learning based framework 2.1 on face scanning pattern analysis as an alternative ASD measurement. Compared with traditional instruments such as ADOS-G and ADOS-2, the proposed framework requires much less human interaction and expertise. We do not argue that such framework can completely replace traditional ones. Rather, it can be regarded as a supplement that ben-

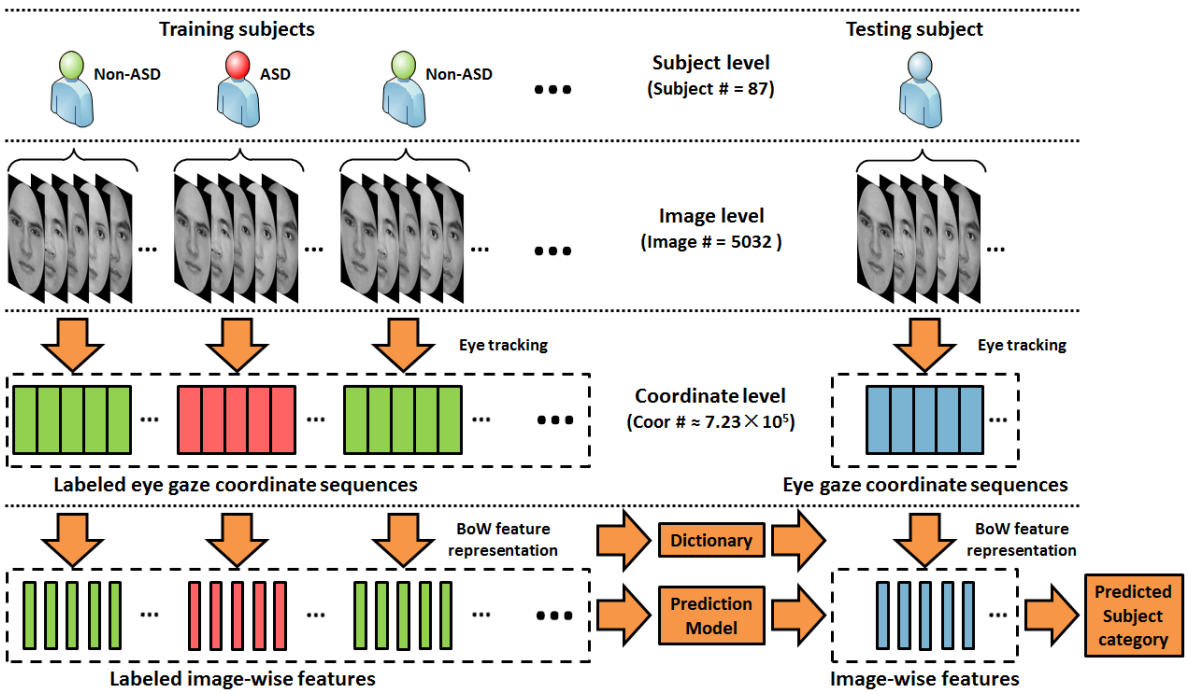

Figure 2.1: An overview of the proposed evaluation protocol. Each subject views a set of face images, while the set of eye gaze coordinates on each viewed image are recorded using eye tracking devices. The proposed method encodes the eye gazes at image level with the BoW model.

efits earlier and more accurate ASD diagnosis. While there are studies that also use machine learning to optimize the diagnosis process, these studies do not change the highly interactive essence of traditional diagnosis procedure. Different from previous literature focusing on the statistical significance of ASD symptoms conveyed by the face scanning patterns, we address the prediction problem and seek to propose a machine learning solution to measure the potential ASD risk. Particularly, we use a data-driven approach to extract features from the face scanning data and support vector machine (SVM) for classification.

2.2 Dataset

We consider two datasets in this chapter. The first one is the Child Dataset proposed in [31]. The child dataset used in the current paper included three groups of participants: 29 4- to 11-year-old Chinese children with ASD, 29 Chinese TD children matched with the chronological age, and another group of 29 Chinese TD children matched with IQ (see Table 1 for details). All children with ASD were diagnosed by experienced clinicians and met the diagnostic criteria for Autism Spectrum Disorder according to the

DSM-IV. Due to the limited access to the ADI-R and ADOS in China, we confirmed the diagnosis using the Chinese version of Autism Spectrum Quotient: Children's Version. Children were asked to memorize six faces (three Chinese faces as the Asian faces and three Caucasian faces), and later tested to recognize these faces from 18 novel faces, including Asian and Caucasian faces (width: 500 pixels, height: 700 pixels, resolution: 72 pixels per in.). All face stimuli were gray-scale and front-view, with their external features (e.g. hair and ears) removed with an ellipse shaped window. There were there study blocks, and in each study block, children were asked to remember two faces (one Chinese face and one Caucasian face, each presented for 3 sec). Each study block was followed by three test blocks in which children were asked to identify whether each face was seen before or not. Each test block comprised two target faces and two foil faces, which were presented until children responded. Children's eye movements during the study and test phases were recorded by a Tobii T60 eye-tracker (sample rate: 60 Hz; both eyes were tracked) with the Tobii Studio software. More details of the participants, the material, and the experimental procedures were provided in the Yi et al..Yi et al. analyzed the eye movement data based on the area of interest (AOI) approach to compare the fixation duration between groups within each predefined face region (e.g. eyes, nose, and mouth).

The Adult Dataset focuses on adolescents and young adults, and is a slightly cleaned up version of the dataset used in [78, 79], including 19 ASD and 46 non-ASD young adults[1]. On top of the above two datasets, we propose a machine learning framework to analyze the eye movement data during face processing so as to identify the ASD symptoms.

2.2.1 Notations and definitions

Before delving into technical details, we list the notations and their definitions in section for the algorithmic clarity. For the rest of the chapter, we use $\boldsymbol{X} = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^2, i \in 1, ..., N\}$ to denote the entire set of 2D eye fixation coordinates on all the viewed faces from all the participants in the training set, where N is the total number of coordinate samples. We use $\boldsymbol{X}^+ = \{\boldsymbol{x}_i | i \in 1, ..., N^+\}$ to denote the set of coordinates from the participants diagnosed with ASD in the training set, and similarly $\boldsymbol{X}^- = \{\boldsymbol{x}_i | i \in 1, ..., N^-\}$ the set of

[1]As a result, the results between this paper and [79] may have certain mismatch, and are not directly comparable.

coordinates from the rest participants, where $\boldsymbol{X} = \boldsymbol{X}^+ \cup \boldsymbol{X}^-$.

2.3 Feature Representation: Codebook Learning

2.3.1 Bag of words feature representation

Feature representation is a crucial part of our classification framework to select relevant features for the classification purpose. A discriminative (good) feature should maximally reveal the statistical difference between participants from different groups, while being minimally sensitive to intra-group variations. While the sequence of eye fixation coordinates or face regions can be incorporated as a temporal feature, we did not adopt these temporal features here due to the sparseness of cross-region transition in our training data. We therefore considered order-less features where the measurement is not sensitive to temporal order of coordinates. What we measured was the frequency distribution of coordinates which treated all face scanning coordinates equally without temporal information. More importantly, we used the frequency distribution as a discriminative feature for the ASD classification, considering the existing evidence on the correlation between the coordinate frequency distribution when scanning faces and the ASD symptoms. Numerous studies have indicated that children and adults with ASD show atypical visual attentions to faces compared to their TD counterparts. Such a face scanning atypicality was directly reflected in the abnormality of ASD in the distributions of fixation coordinates, which serves as an feature in our framework. The feature representation includes two procedures: the facial region partitioning with k-means and the histogram feature extraction.We performed the quantization of fixation coordinates with the k-means algorithm, where fixation coordinates are clustered and divided into K different clusters with distinct cluster centroids, as shown in Figure 2. The k-means quantization was conducted based on the fixation coordinate data of all participants. Each observed coordinate was assigned to the cluster with the closest centroid. Such quantization results in the partitioning of face images into K different cell-like regions, such that fixations falling into the same region indicate close proximity in the visual attention location. Compared to the well-known Area of Interest (AOI) based approach, our quantization strategy was more data-driven oriented. The AOI approach is a top-down process which determines the partitioned region boundary empirically and could be influenced by the semantic meaning of face parsing

without statistical justification. In contrast, our data-driven approach can represent face scanning hot spots by generating partitions based on statistical distribution of coordinates.Given the sequence of the fixation coordinates from each face viewed by every participant, we assigned each coordinate to the most proximal cluster centroid obtained by k-means and counted the number of assignments for each cluster. Then the assignment counts were normalized by being divided by the total number of coordinates. As a result, a histogram feature was defined to decode the frequency distribution of the visual attentions on each part of the face. Since the extracted histogram is an image-level feature encoding the visual attention on a single face, we repeated the histogram extraction process for every face viewed by each participant, to obtain a training set whose labels are determined by the participant categories (i.e. ASD group, aged-matched TD group, and IQ-matched TD group).

The feature we consider is the BoW (Bag of words) histogram representation on the gaze coordinates. The BoW model originally came from the linguistic community [80] and has ever since been a very popular feature representation framework with wide applications in Natural Language Processing, information retrieval [81] and computer vision [82]. The reason why such model is called "bag-of-words" is because a sentence or a document can be represented as the bag (multi-set) of its words, disregarding grammar and even word order but keeping multiplicity.

Similar analogies can be made here as we treat the centers of concentrated visual attentions as dictionary words, while the sequence of coordinates per image per subject as one document. An atypical frequency distribution of gaze on different parts of a face image can be a strong evidence of reduced visual attention. Motivated by the abnormal visual attentions observed in many ASD studies, we proposed an ASD children identification framework with BoW, k-means dictionary learning and SVM, and are so far the most related studies to this work.

2.3.2 Dictionary learning: AOI and k-means

One of the most important problem in of BoW feature representation is how to obtain the dictionary word. We will discuss the dictionary learning in this part.

AOI aims to measure eye gazes that fall within a predefined areas of interest, typically including

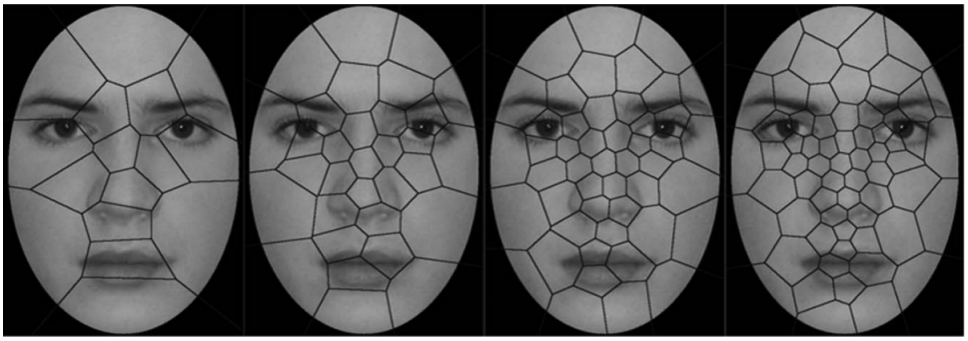

Figure 2.2: Illustration of partitioned face regions by k-means with different cluster numbers (K).

eyes, nose and mouth. With the defined AOIs, one is able to statistically estimate the frequency counts of eye fixations on different face areas. A common problem with the AOI approach is that it tends to lump fixations to a relatively large area without further discrimination. The boundary of AOI is often determined empirically and is influenced by the semantic meaning of face parsing without statistical and psychological justification, while a subject's visual attention could in fact be largely influenced by certain sub-AOI regions highly responsive to human brains as mid-level visual features.

To discover important spatial regions for eye movement patterns, we use the k-means algorithm to cluster the recorded eye gaze coordinates from all participants in the training set, and divide the face image into different sub-regions. The resulting output is a set of cell-like spatially partitioned face regions indicating clusters of gazes that are relatively more concentrated. Four partitioning examples with dictionary numbers respectively equal to 16, 32, 48 and 64 are shown in Fig. 2.2.

The difference between AOI and k-means is that the latter is data-driven which does not require manual partitions (See the right image in Fig. 2.3).

2.4 Learning Discriminative BoW Dictionary

Dictionary learning presents an important problem in BoW representation as the quality of learned codebook has direct impacts on the quality of represented features. Often, one would hope that the dictionary can encode as much discriminative information as possible, such that the feature coefficients on this dictionary show significant inter-class differences which benefit the classification task. An important question one may ask is: How to quantitatively measure the quality of the given codebook?

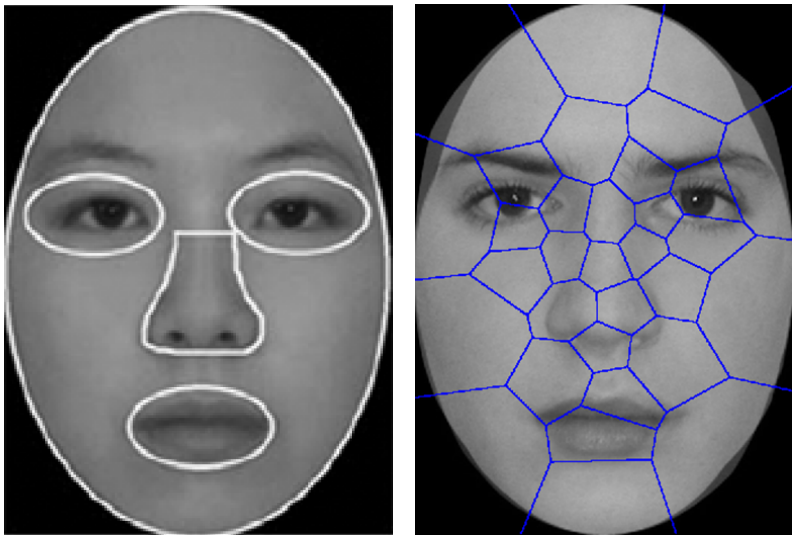

Figure 2.3: Illustration of the dictionary words projected onto the viewed image. Left: Partitioned regions as dictionary words in AOI. Right: Regions learned by k-means.

2.4.1 A unified view towards codebook quality

In [64], the authors proposed the concept of **purity** to measure how discriminative a dictionary word is, and **coverage** to measure how representative it is. We follow this idea to learn dictionaries that have larger values in both terms. Given a certain cluster partition $\boldsymbol{C}$ and the cluster index k, the purity for positive class $P^+(k \mid \boldsymbol{C})$ can be modeled as:

$$P^+(k \mid \boldsymbol{C}) = \frac{N^+(k \mid \boldsymbol{C})}{N^+(k \mid \boldsymbol{C}) + N^-(k \mid \boldsymbol{C})}, \tag{2.1}$$

where $N^+(k \mid \boldsymbol{C})$ and $N^-(k \mid \boldsymbol{C})$ respectively denotes the numbers of positive and negative samples assigned to cluster partition C. Again, note that such measurement differs from the density ratio in [64], in the sense that Eq. (2.1) is normalized between 0 and 1. Similarly, the purity of negative class can be defined as:

$$P^-(k \mid \boldsymbol{C}) = \frac{N^-(k \mid \boldsymbol{C})}{N^+(k \mid \boldsymbol{C}) + N^-(k \mid \boldsymbol{C})}. \tag{2.2}$$

While it is desirable to increase the dictionary purities for both classes, increasing the purity of both positive and negative classes is contradicting in the same word. What truly matters is the difference of sample numbers and its ratio versus the word size. As a result, we look into the following purity measure:

$$P(k \mid \boldsymbol{C}) = \frac{|N^+(k \mid \boldsymbol{C}) - N^-(k \mid \boldsymbol{C})|}{N^+(k \mid \boldsymbol{C}) + N^-(k \mid \boldsymbol{C})}, \tag{2.3}$$

which is able to measure the level of purity for both classes with a unified representation. On the other hand, the coverage for positive and negative class can be modeled as:

$$C(k \mid \boldsymbol{C}) = N^+(k \mid \boldsymbol{C}) + N^-(k \mid \boldsymbol{C}). \tag{2.4}$$

A dictionary ideally should have good purity and coverage simultaneously. A natural way is to treat the product of both benchmarks as the objective, which shares similar spirit to the f-measure[2]. Therefore, the word quality can be estimated as:

$$Q(k, \boldsymbol{C}) \triangleq P(k \mid \boldsymbol{C})C(k \mid \boldsymbol{C}) = |N^+(k \mid \boldsymbol{C}) - N^-(k \mid \boldsymbol{C})| \tag{2.5}$$

The problem of finding a good dictionary word can therefore be formulated as maximizing the quality estimation objective with respect to k and $\boldsymbol{C}$:

$$\max_{\boldsymbol{C}} \sum_{k} Q(k, \boldsymbol{C}) \tag{2.6}$$

2.4.2 Approximating with kernel density estimation

Directly optimizing the objective in Eq. (2.6) is difficult since the optimization is non-continuous, non-convex, and the solution space of $\boldsymbol{C}$ is huge. Our approach here is to approximate with kernel density estimation and mode-seeking. Specifically, when the size of each dictionary word is reasonably small, a good approximation to $N^+(k \mid \boldsymbol{C})$ and $N^-(k \mid \boldsymbol{C})$ is the local density estimator:

$$\hat{P}(\boldsymbol{x}_k \mid \boldsymbol{X}^+) \propto N^+(k \mid \boldsymbol{C}) \qquad \hat{P}(\boldsymbol{x}_k \mid \boldsymbol{X}^-) \propto N^-(k \mid \boldsymbol{C}) \tag{2.7}$$

where $\boldsymbol{x}_k$ is the location of the k-th dictionary word in feature space. In addition, we define $\hat{P}(\boldsymbol{x}_k|\boldsymbol{X}^+)$ to be the following Gaussian kernel density estimator:

$$\hat{P}(\boldsymbol{x} \mid \boldsymbol{X}^+) \triangleq \frac{c_d}{Nh^d} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}^+} \exp\Big(-\frac{||\boldsymbol{x} - \boldsymbol{x}_i||^2}{2h^2}\Big) \qquad \hat{P}(\boldsymbol{x} \mid \boldsymbol{X}^-) \triangleq \frac{c_d}{Nh^d} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}^-} \exp\Big(-\frac{||\boldsymbol{x} - \boldsymbol{x}_i||^2}{2h^2}\Big) \tag{2.8}$$

where $d = 2$ is the dimension, h is the bandwidth that controls the kernel smoothness, and $c_d = 2\pi^{(-d/2)}$ is a normalization constant. The word quality located at $\boldsymbol{x}$ can thus be estimated as:

$$Q(\boldsymbol{x}) = |\hat{P}(\boldsymbol{x} \mid \boldsymbol{X}^+) - \hat{P}(\boldsymbol{x} \mid \boldsymbol{X}^-)| \tag{2.9}$$

[2]One may also consider linear combination but this leads to the weight issue between purity and coverage for the different scales.

2.4.3 Finding $Q(x)$ local maxima with dual mode seeking

Our goal is to find a set of local maxima of $Q(x)$ which indicate the locations of high quality words. Note that Eq. (2.9) is a continuous function with respect to $\boldsymbol{x}$. This allows one to optimize it with respect to $\boldsymbol{x}$ using gradient ascent. Since Eq. (2.9) contains absolute values, we consider the alternative objective:

$$Q^*(\boldsymbol{x}) = \hat{P}(\boldsymbol{x} \mid \boldsymbol{X}^+) - \hat{P}(\boldsymbol{x} \mid \boldsymbol{X}^-) \tag{2.10}$$

Assuming that the gradient ascent/descent process guarantees the monotonic increasing/decreasing of $Q^*(\boldsymbol{x})$, we have the following theorems:

Proposition 1: $Q(\boldsymbol{x}) = -Q^*(\boldsymbol{x}), \forall \boldsymbol{x} \in \{\boldsymbol{x} | Q^*(\boldsymbol{x}) < 0\}$.

Remark: The proof is omitted as it is strightforward. Proposition 1 indicates that the landscape of $Q(\boldsymbol{x})$ is equal to flipping the negative part of $Q^*(\boldsymbol{x})$ as positive.

Proposition 2: Gradient ascent on $Q(\boldsymbol{x})$ is equal to gradient ascent on $Q^*(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \{\boldsymbol{x} | Q^*(\boldsymbol{x}) > 0\}$.

Proposition 3: Gradient ascent on $Q(\boldsymbol{x})$ is equal to gradient descent on $Q^*(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \{\boldsymbol{x} | Q^*(\boldsymbol{x}) < 0\}$.

Remark: Proposition 2 and 3 can be directly concluded from Proposition 1. As a result, performing mode seeking on $Q(\boldsymbol{x})$ can be alternatively done by performing dual gradient ascent/descent on $Q^*(\boldsymbol{x})$ with respect to the gradient $\nabla Q^*(\boldsymbol{x})$. To simplify the computation, note that we have:

$$\nabla \hat{P}(x \mid \boldsymbol{X}^+) = \frac{1}{h^2} \hat{P}(x \mid \boldsymbol{X}^+)(\boldsymbol{x}_{m^+} - \boldsymbol{x}) \qquad \nabla \hat{P}(x \mid \boldsymbol{X}^-) = \frac{1}{h^2} \hat{P}(x \mid \boldsymbol{X}^-)(\boldsymbol{x}_{m^-} - \boldsymbol{x}) \tag{2.11}$$

where $\boldsymbol{x}_{m^+}$ is the weighted mean of positive data samples weighted by kernels:

$$\boldsymbol{x}_{m^+} = \frac{\sum_{\boldsymbol{x} \in \boldsymbol{X}^+} \exp(-||\boldsymbol{x} - \boldsymbol{x}_i||^2 / 2h^2) \boldsymbol{x}_i}{\sum_{\boldsymbol{x} \in \boldsymbol{X}^+} \exp(-||\boldsymbol{x} - \boldsymbol{x}_i||^2 / 2h^2)} \tag{2.12}$$

$\boldsymbol{x}_{m^-}$ is defined similarly. The gradient of objective function is therefore computed as:

$$\nabla Q^*(\boldsymbol{x}) = \frac{1}{h^2} \left[\hat{P}(\boldsymbol{x} | \boldsymbol{X}^+)(\boldsymbol{x}_{m^+} - \boldsymbol{x}) - \hat{P}(\boldsymbol{x} \mid \boldsymbol{X}^-)(\boldsymbol{x}_{m^-} - \boldsymbol{x}) \right] \tag{2.13}$$

One could see that Eq. (2.13) is basically a weighted combination of the mean shift vectors [83] from positive and negative samples, where the weights are the kernel densities. Accordingly, one may consider the following dual mode seeking step to find local maxima of $Q(\boldsymbol{x})$, see Algorithm 1:

Algorithm 1 Dual mode seeking

Estimate word quality located at location $\boldsymbol{x}$: $\hat{P}(\boldsymbol{x}_i|\boldsymbol{X}^+) - \hat{P}(\boldsymbol{x}_i|\boldsymbol{X}^-)$
while not converged **do**
  if $\hat{P}(\boldsymbol{x}_i|\boldsymbol{X}^+) - \hat{P}(\boldsymbol{x}_i|\boldsymbol{X}^-) > 0$ **then**
    perform mode seeking with $\nabla Q^*(\boldsymbol{x}_i)$ until convergence (gradient ascent)
  else if $\hat{P}(\boldsymbol{x}_i|\boldsymbol{X}^+) - \hat{P}(\boldsymbol{x}_i|\boldsymbol{X}^-) < 0$ **then**
    perform mode seeking with $-\nabla Q^*(\boldsymbol{x}_i)$ until convergence (gradient descent)
  end if
end while

2.4.4 Dual mode seeking as supervised mean shift

In reality, one does not need to explicitly flip the sign of $\nabla Q^*(\boldsymbol{x})$ in order to perform dual mode seeking. Let $y_i \in \{1, -1\}$ indicates the label of $\boldsymbol{x}_i$, the Eq. 2.13 can be re-written as:

$$\nabla Q^*(\boldsymbol{x}) = \frac{c_d}{Nh^{d+2}} \Big[\sum_{i=1}^{N} y_i k(\boldsymbol{x}, \boldsymbol{x}_i) \Big] \Big[\frac{\sum_{i=1}^{N} y_i k(\boldsymbol{x}, \boldsymbol{x}_i) \boldsymbol{x}_i}{\sum_{i=1}^{n} y_i k(\boldsymbol{x}, \boldsymbol{x}_i)} - \boldsymbol{x} \Big] \tag{2.14}$$

where we have:

$$\sum_{i=1}^{N} y_i k(\boldsymbol{x}, \boldsymbol{x}_i) = \sum_{i=1}^{N} y_i \exp\Big(- \frac{||\boldsymbol{x} - \boldsymbol{x}_i||^2}{2h^2} \Big) = \frac{Nh^d}{c_d} Q^*(\boldsymbol{x}) \tag{2.15}$$

Note that dividing Eq. 2.14 with $\sum_{i=1}^{n} y_i k(\boldsymbol{x}, \boldsymbol{x}_i)$ actually gives a generalized form of mean shift. Also, the sign of $g(\boldsymbol{x})$ is exactly determined by $\sum_{i=1}^{n} y_i k(\boldsymbol{x}, \boldsymbol{x}_i)$. One may cancel the flipping sign of dual mode seeking simply by iteratively shifting with the following mean shift vector:

$$\boldsymbol{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} y_i k(\boldsymbol{x}, \boldsymbol{x}_i) \boldsymbol{x}_i}{\sum_{i=1}^{n} y_i k(\boldsymbol{x}, \boldsymbol{x}_i)} - \boldsymbol{x} \tag{2.16}$$

Note that an interesting aspect of the above mode seeking algorithm (Eq. (2.14) - Eq. (2.16)) is that it can be viewed as a generalized form of supervised mean shift algorithm, where the labels y_i introduce class-aware discriminative information into the learning process.

2.5 Summary and Implementation Details

Zooming out a bit, we briefly recap our full picture. We started from the motivation to capture local modes that maximize the difference between ASD and non-ASD subjects on the attention maps. Our goal is to automatically identify these modes through a data-driven method in contrast to manual selection. In Section 2.4.1, we start by defining quantitative measures of the dictionary (cluster) quality with purity

and coverage. We then define the dictionary quality as the multiplication of purity and coverage. We approximate the dictionary quality with kernel density estimation in Section 2.4.2, and further approximate the optimization of dictionary as dual mode seeking in Section 2.4.3. Finally, we show that the proposed dual mode seeking method can be generalized into a supervised mean shift form in Section 2.4.4, and addresses convergence issues in Section 2.5.1.

2.5.1 Convergence with back tracking line search

Unfortunately, unlike the conventional mean shift, performing gradient ascent with Eq. (2.16) does not guarantee the monotonic increase of gradient and algorithm convergence, since the sum of kernel weights contains negative terms. This often happens when the densities of positive and negative classes are approximately equal to each other. In this case the denominator of Eq. (2.16) is very small, leading to relatively large shifting vector or potential numerical issues. This can be practically solved by adaptive step size normalization with respect to the denominator and step size reduction with back tracking line search. Whenever the quality objective value of the next step is not increased, back tracking line search multiplies the current step size with 0.5. This guarantees the monotonic increase of the objective and the algorithm convergence. In practice we observe that mean shift with Eq. (2.16) works well at most feature space positions, and the need for performing back tracking line search is reduced very fast as the density of one class quickly dominates over another.

2.5.2 From discriminative modes to BoW representation

The supervised mean shift algorithm in Sec. 2.4.4 returns a set of local maxima of $Q(x)$ which indicate locations of high quality dictionary words. The subsequent question is how to transform these maxima into BoW representation by learning a particular clustering configuration $\boldsymbol{C}$ that favors these locations.

To this end, we consider a mean shift based clustering method to obtain the dictionary words and $\boldsymbol{C}$. The idea here is to initialize a set of kernel locations $\boldsymbol{x}$ with the coordinate samples and iteratively apply supervised mean shift in Sec. 2.4.4 to each kernel for adequate number of iterations. This will basically shift each of the kernel from its initial feature space location to local maxima of $Q(x)$ through gradient

ascent. We then treat these shifted kernels as data samples and use k-means to obtain a total of K cluster centroids which are mostly located on the $Q(x)$ maxima.

Specifically, we use all the fixation coordinate samples in the training set for density estimation. For speed purpose, we sample 1 out of 20 training coordinates to initialize the kernel locations, and perform 30 rounds of mean shifts on these kernels. We keep these settings the same across all our experiments. Once obtaining the cluster centroids, we assign each coordinate sample to the nearest centroid, therefore obtaining a cluster labeling $\boldsymbol{C}$ and the dictionary words. We then use the words to compute the BoW feature to encode the fixation coordinates for each sequence.

2.6 Classification

The classification is the process to use the selected features to assign group labels to the participants. Given the labeled features, the classification algorithm could build a classifier that assigns new examples into different categories. The classification process included the following steps: the generation of the training and testing data, the image-level classification, and the participant-level classification. Given a set of labeled participants, we used the leave-one-out cross-validation strategy to separate the original data into the training dataset and the testing dataset. Each time, one out of all participants was selected sequentially as the testing participant while the classification model was learned according to the histogram features from the rest of the participants. The learned model was then tested, and then both the image-level scores and the participant-level scores for the test participant were returned. Such a leave-out-and-evaluate process was repeatedly performed for all participants included in our dataset.

We started with the image-level ASD classification based on the extracted histograms containing the visual attention information on single faces, followed by the participant-level classification. At the training stage, a SVM classifier was trained based on the labeled histograms. The SVM attempted to find a linear decision boundary with a maximum margin separating the data into two classes. Considering that the data were not linearly separable, we adopted the Radial Basis Function (RBF) kernel SVM which performed a nonlinear projection of the data into a high-dimensional space to make the data more linearly separable.

Image-level predictions may not be robust enough due to the limitation of information conveyed by

every test on a single image. A subject level prediction on the other hand is what we ultimately desired. Therefore, we ensemble image-level predictions to finalize subject-level predictions by fixing a global threshold. We consider the following two ways of ensemble strategies:

Soft prediction score: The RBF kernel SVM gives each image-level testing feature a soft prediction score. Suppose each feature is re-denoted as $\boldsymbol{h}_n$ with a single, global index. Also let $y_n \in \{-1, 1\}$ denotes the label of feature $\boldsymbol{h}_n$, $\boldsymbol{w}$ and b denote the learned parameters defining the decision hyperplane. The soft prediction score is computed as:

$$S_{\text{soft}}(n) = \boldsymbol{w}^\top \boldsymbol{\Phi}(\boldsymbol{h}_n) + b = \sum_m \alpha_m y_m K(\boldsymbol{h}_m, \boldsymbol{h}_n) + b. \tag{2.17}$$

The top row of (2.17) gives an intuitive geometric interpretation of the score. It is basically the functional margin of a kernelized feature and the decision boundary. In practice however, the prediction score is obtained by solving the dual problem of SVM, resulting in the second row where α_i are the introduced lagrange multipliers and $K(\cdot, \cdot)$ is the kernel function. The subject-level mean score is defined as:

$$S_{\text{sub}}(i) = \frac{1}{|S_i|} \sum_{n \in S_i} S_{\text{soft}}(n), \tag{2.18}$$

where S_i corresponds to the set of global indexes belonging to the i-th subject.

Hard prediction score: The RBF kernel SVM gives each image-level testing feature with a $\{0, 1\}$ hard score:

$$S_{\text{hard}}(n) = \begin{cases} 1, & \text{if} \quad S_{\text{soft}}(n) > 0 \\ 0, & \text{else} \end{cases} \tag{2.19}$$

Again the subject-level mean score is defined as:

$$S_{\text{sub}}(i) = \frac{1}{|S_i|} \sum_{n \in S_i} S_{\text{hard}}(n), \tag{2.20}$$

The subject level prediction for both methods is determined with a global threshold T:

$$S_{\text{sub}}(i) \underset{non-ASD}{\overset{ASD}{\gtrless}} T \tag{2.21}$$

During the testing phase, the learned SVM model made a classification for the group membership using each testing histogram feature with a corresponding confidence score. The sign of the score can be either positive or negative to indicate the classification of the histogram feature. The absolute value of

the confidence score measures the distance of the testing sample from the decision boundary. A higher confidence score indicates a more confident classification decision. The image-level ASD classification indicates the likelihood of the ASD symptoms only based on the face scanning patterns from a single face. It was more meaningful to make the classification decision at the participant level to indicate the likelihood of the ASD symptoms for each participant. We therefore defined the participant-level classification score as the average image-level classification score of each participant, as shown in Figure 3. Considering that the imbalanced ASD and TD training set sizes may cause biased SVM classifications, we introduced a flexible threshold instead of zero to determine the final classification labels. The participants with the classification score above the threshold were labeled as individuals potentially with ASD, while those ones whose classification score below the threshold were labeled as TD individuals.

2.7 Experimental Results

In this section, we report comprehensive evaluations of our method on the two datasets described in Section 2.2. Note that the Adult Dataset is a slightly cleaned up version of the dataset used in [78] and [79]. As a result, the results on the adult dataset between this chapter and [79] may have certain mismatches, and are not directly comparable.

2.7.1 Evaluation protocol

Following [79] and [84], we evaluate the proposed method by leave-one-out cross-validation testing, where each subject is consecutively held out for testing while the rest are used for training. By doing this each time we divide the image-level BoW features into two sets: one for testing and the other for training a prediction model. Following [79, 84], we train an RBF kernel SVM as the prediction model, and predict the test subject score as the mean over the soft SVM prediction scores on the images viewed by each test subject. Finally, a global threshold T is set for all testing subjects to obtain the subject-level predictions. For the fairness of comparison, we vary and search the hyperparameters of all comparing methods and report the best performance. Specifically, For the proposed method and baselines which include the k-means clustering step, we search the number of clusters within {35, 40, 45, 50, 55, 60, 65, 70}. We also

search the γ and C values in kernel SVM for all comparing methods, by varying them as exponentials of 2. The search ranges of γ and C are set to $2^{-6} \sim 2^{0}$ and $2^{6} \sim 2^{16}$, respectively.

2.7.2 Evaluation benchmarks

We consider the following benchmarks to quantitatively evaluate the prediction performance:

Accuracy (Acc): The number of correctly predicted subjects versus the total number of subjects.

Area under the curve (AUC): The total area under the ROC curve versus the whole area. And the ROC curve is a set of (subject-level) true positive rates versus false positive rates obtained by synchronously varied the global threshold T for all testing predictions.

Purity: To analyze the level of determinativeness of the dictionaries learned by different methods, we also visualize the dictionary purity profile of comparing methods.

Sensitivity: Ratio of correct true positives versus positives.

Specificity: Ratio of correct true negatives versus negatives.

2.7.3 Baselines

We compare with several dictionary learning baselines that are closely related to BoW representations:

K-means. Using k-means for dictionary learning as described and reported in [84].

Class K-means. Performing k-means on both positive and negative data separately with approximately the same number of clusters.

Mean Shift. Applying the conventional mean shift [83] approach on all the data, followed by k-means dictionary learning.

Class Mean Shift. Applying the conventional mean shift approach on both positive and negative data separately, followed by k-means dictionary learning.

Disc Mode Seek. Applying discriminative mode seeking [64] on all the data, followed by k-means dictionary learning.

Note that both class k-means and class mean shift can be regarded as variants of [60] where the concept of class-aware BoW representations is adopted to our problem.In addition, the bandwidths of density

estimators in mean shift, class mean shift and the proposed method are also cross-validated.

2.7.4 Main results on Child Dataset

Follwing [84], we comprehensively evaluate the proposed method and baselines on the complete child dataset as well as controlled scenarios where the non-ASD group is divided into IQ-matched and age-matched groups. We denote these two settings as "ASD vs. TD-IQ" and "ASD vs. TD-Age", respectively. Results of the comparing methods are reported in Table 2.1, indicating that the proposed overall performs better.

For child dataset, each child is shown with face images from two sources: Asian Faces and Caucasian Faces. This is another typical setting in psychology study to analyze the ASD behavior. Following this setting, we subdivide our dataset into two subsets, and conduct the same evaluation. Table 2.2 shows the results of the proposed method and comparing baselines on the child dataset. One could see that compared with other baselines, our method has the highest accuracy (91.95%) and AUC (93.4%) on the full dataset as well as on the Asian Faces and Caucasian Faces subsets. This shows the benefit from the improved dictionary word quality using our method.

Table 2.1: Main results on child dataset with different TD Groups, where "TD-IQ" indicates "IQ-matched group" and "TD-Age" indicates "age-matched group".

Dataset	**All Data**		**ASD vs. TD-IQ**		**ASD vs. TD-Age**	
Evaluation Metric	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
K-means [84]	88.51	89.63	86.21	88.94	84.48	85.37
Class K-means	87.36	90.79	83.91	84.74	82.76	85.38
Mean Shift	89.66	92.51	87.93	88.59	**87.93**	**88.59**
Class Mean Shift	88.51	92.83	86.21	89.08	86.21	86.87
Disc Mode Seek [64]	89.66	92.64	87.93	89.34	**87.93**	87.03
Proposed	**91.95**	**93.40**	**89.66**	**90.96**	**87.93**	87.45

Table 2.2: Main results on child dataset with different face subsets. Asian Faces and Caucasian Faces indicates that viewed faces type.

Dataset	**All Data**		**Asian Faces**		**Caucasian Faces**	
Evaluation Metric	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
K-means [84]	88.51	89.63	81.61	82.40	**90.80**	94.41
Class K-means	87.36	90.79	86.21	84.13	89.66	93.40
Mean Shift	89.66	92.51	85.06	**86.50**	90.80	93.34
Class Mean Shift	88.51	92.83	85.06	84.58	89.66	93.87
Disc Mode Seek [64]	89.66	92.64	85.06	85.34	89.66	94.03
Proposed	**91.95**	**93.40**	**87.35**	86.27	**90.80**	**94.48**

Table 2.3: Main results on Adult Dataset.

	Method	Accuracy	AUC
Adult Dataset	K-means [84]	72.31	71.51
	Class K-means	73.85	66.48
	Mean Shift	72.31	68.97
	Class Mean Shift	73.85	72.77
	Disc Mode Seek [64]	**75.39**	73.37
	Proposed	**75.39**	**75.06**

2.7.5 Main result on Adult Dataset

Following the experimental settings of the complete Child Dataset, we also evaluate the proposed method and baselines on Adult Dataset, with the results reported in Table 2.3. One could again observe that our method outperforms all comparing baselines with a sizable margin.

2.7.6 ROC Curves

We show the ROC curves of all the comparing methods on both the child dataset and the adult dataset in Figure 2.4. In general, an ROC curve closer to top left corner indicates the better prediction quality of a model. This can be quantified by the AUC score, an better reflection of the holistic ROC curve performance than accuracy since AUC is a cumulative measure over the entire range of thresholds. Overall, one could see that our method (in blue color) gives the best performance in the ASD and non-ASD classification task. The corresponding AUC scores are shown in both Table 2.2 and Table 2.3. The results show that the AUC scores are 93.4% on the child dataset and 75.06% on the adult dataset.

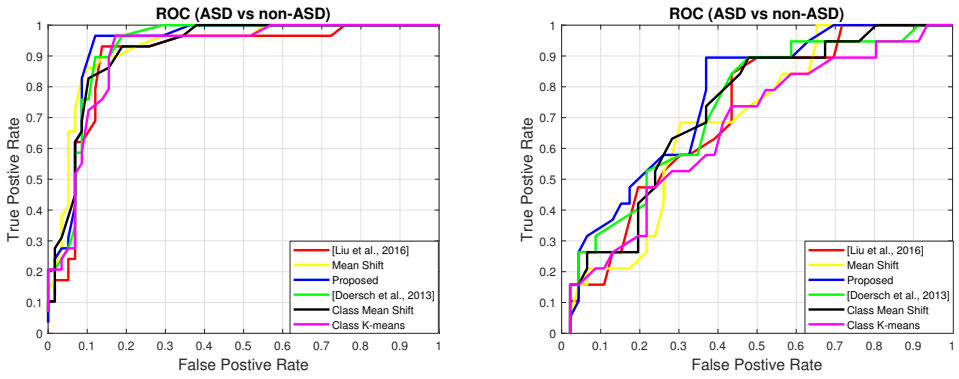

Figure 2.4: ROC Curves of all comparing methods. Left: Child Dataset. Right: Adult Dataset. Best viewed in color.

2.7.7 Sensitivity and specificity

We report the sensitivity and specificity scores of all comparing methods in Table 2.4, where the proposed method overall outperforms comparing methods on both child data with $Sensitivity = 0.966$ and on adult data $Sensitivity = 0.316$. The high sensitivity means that our proposed method have few false negative results, and thus fewer cases of disease are missed. The sensitivity is very important for effective screening program. And result shows that proposed machine learning method would be useful for ASD early screening. We also discuss the performance difference on child and adult dataset in Section 2.7.11.

Table 2.4: Sensitivity and specificity scores on child dataset and adult dataset.

Dataset	**Child Dataset**		**Adult Dataset**	
Eval Metric	Sensitivity	Specificity	Sensitivity	Specificity
K-means [84]	0.931	0.862	0.158	0.957
Class K-means	0.966	0.897	0.158	**0.978**
Mean Shift	0.862	**0.914**	0.211	0.934
Class Mean Shift	0.828	**0.914**	0.263	0.934
Disc Mode Seek [64]	0.897	0.897	0.263	0.957
Proposed	**0.966**	0.897	**0.316**	0.934

2.7.8 Sensitivity to SVM parameters

Although slightly different optimal configurations may apply for different methods, we observe a general trend that all comparing methods tend to work best around $\gamma = 2^{-3} \sim 2^{-4}$ and $C = 2^{13} \sim 2^{14}$. We also observe a clear pattern for every method that similar top results appear with multiple combinations of $\gamma - C$ pairs: increasing γ requires decreased C. Most importantly, all comparing methods are not sensitive to the parameters - usually with a universal $1 \sim 2\%$ decrease of performance within a large parameter range.

2.7.9 Dictionary purity analysis

To analyze the discriminativeness of the dictionaries learned by different methods, we also compare the word purities of different methods in Figure 2.5. In particular, we first sort the dictionary words from high to low by the positive class purity, and then plot the purity of the top ranked words. One could see from Figure 2.5 that the proposed discriminative mode seeking method tends to have higher purities on than

the others. This shows a clear evidence that the proposed method is able to explore the desterminative during dictionary learning.

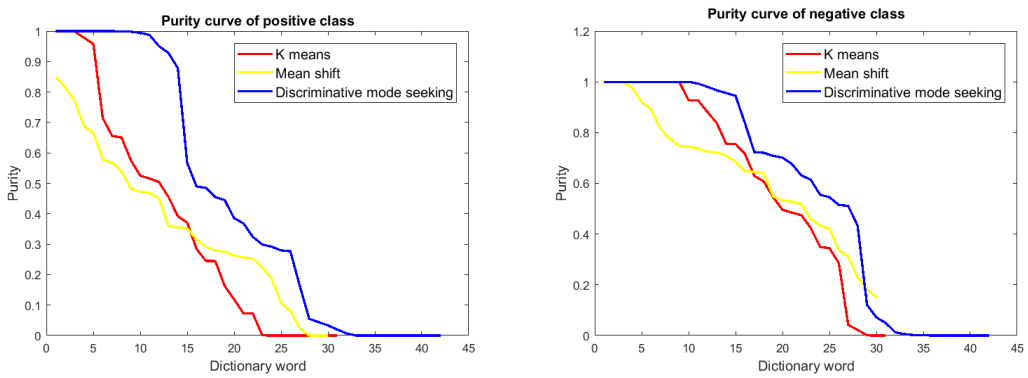

Figure 2.5: Positive and negative purity of different dictionary learning methods on child dataset. Left: Purity curves of positive class. Right: Purity curves of the negative class. Best viewed in color.

2.7.10 Mode seeking visualization

To show how the proposed dual mode seeking works, we visualize the shifted samples at different iterations and compare with mean shift in Figure 2.6. For the results of dual mode seeking, samples with red color indicates that their initial location before shifting belongs to the positive domain, while samples with blue color indicates the opposite. One could see that dual mode seeking is able to correctly find both the positive modes and the negative modes belonging to different classes. However, mean shift tends to find regions with densest samples without considering discriminative class information. It is also very interesting to see that on child Dataset, samples with higher density of negative class tend to concentrate near eyes and the center of the face, which again verifies the strong tendency of less direct eye contacts with ASD children.

2.7.11 Performance difference between Child Dataset and Adult Dataset

Upon comparing the overall identification accuracies on child dataset (Table 2.1) and adult dataset (Table 2.3), one could observe that the performance on adult dataset is not as good as the performance on child dataset. We suspect that when viewing face images, children's reactions are generally more spontaneous than those of adults. This suspect could also be verified by comparing the visualization of dual mode

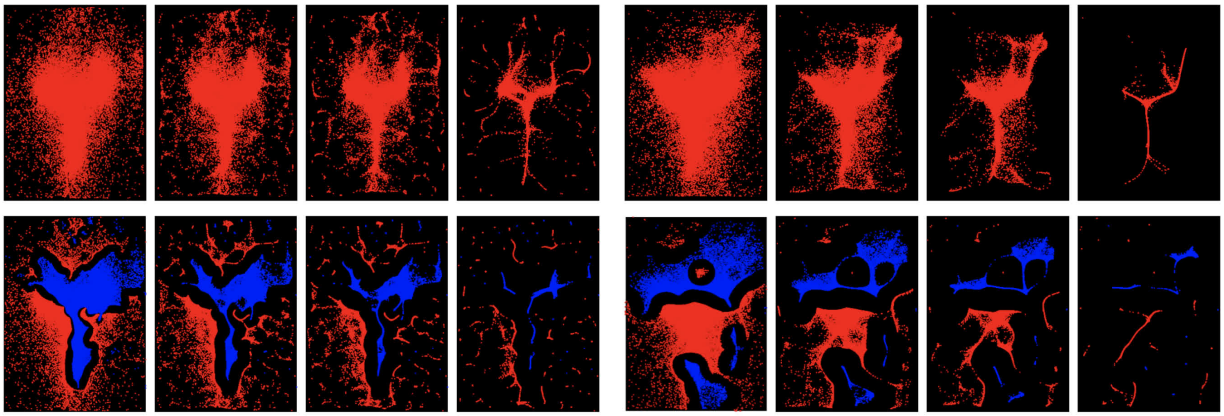

Figure 2.6: Visualization of traditional mean shift (top) and the proposed dual mode seeking (bottom) at different iterations. For dual mode seeking, red indicates $\hat{p}(\boldsymbol{x}_i|\boldsymbol{X}^+) > \hat{p}(\boldsymbol{x}_i|\boldsymbol{X}^-) > 0$, while blue indicates $\hat{p}(\boldsymbol{x}_i|\boldsymbol{X}^+) < \hat{p}(\boldsymbol{x}_i|\boldsymbol{X}^-)$. Left columns: Visualization on child dataset. Right columns: Visualization on adult dataset. Every set of four images correspond to the visualization of shifted samples at iteration 1, 5, 10 and 30 in mean shift or the proposed method. Best viewed in color.

seeking on both child dataset and adult dataset. One could find discriminative regions on the result of child dataset, while the result on adult dataset tends to have less discriminative regions.

2.8 Discussions and Remarks

The Experimental results indicate that our model gives considerable improvement over several widely used dictionary learning methods in terms of representing the face scanning patterns for ASD identification. On the child dataset, our method achieves an accuracy of 91.95% and an AUC score of 93.4%. On different subsets of the child dataset (different TD groups and different face type subsets), our method also outperforms different baselines. The sensitivity and specificity scores of different methods show that our proposed method has the highest sensitivity which may benefit early ASD screening since fewer cases of positive will be missed. However, we notice that the performance on the adult dataset is less promising compared to the child dataset. The conjecture of such observation is stated in Section 2.7.11.

When comparing among the baselines, one could observe a general trend that the the methods based on mode seeking (mean shift, class mean shift, discriminative mode seeking and the proposed method) tend to outperform k-means based method since they generate arbitrary shaped dictionary clusters that better capture important patterns in the feature space. On the other hand, methods based on k-means assume more regular shaped dictionary clusters which are less discriminative. In addition, the connection

between mode seeking based methods and the iMap approach [37] also partly explains the popularity of iMap in the behavioral research community from a pattern recognition perspective.

2.9 Summary

In summary, we propose a novel dictionary learning method based on dual mode seeking. Our method incorporates label information and can automatically mine discriminative dictionary words through supervised mean shift. We also give detailed motivation, intuition, as well as links to psychology studies for the proposed method. Our method can be extended to other types of features as well. For example we could apply the same dictionary learning and BoW representations to motion features and short coordinate sequences in order to incorporate short temporal and higher order information.In addition, the datasets used in this work only contain with children between age 5-10 and adults, with the viewed faces limited to Asian and Caucasian. Including participants with a wider range of ages (especially children), races and genders, together with designing a more comprehensive test protocol, will help to better mitigate the dataset biases and consolidate the psychological discoveries. We will leave this to be addressed and studied in future work.

Chapter 3

Multimodal Children Behavior Dataset Collection

3.1 Motivation

Formal diagnosis of autism often require going through a series of structured and semi-structured examination tasks that involve social interaction between the examiner and the person under assessment. The most widely used protocols include the Autism Diagnostic Observation Schedule-Generic (ADOS-G) and the revised version ADOS-2.The whole process can take up to 90 minutes. As a result, current autism diagnosis often requires considerable amount of labour and cost, which potentially increases the economic burden of ASD families and lowers the chance of early diagnosis.

In light of the above limitations, a major goal of this experiment is to establish a convenient, children-friendly assessment procedure, as well as an intelligent system that can identify signs of autism at a relatively early stage with less or even without professional expertise. Specifically, we believe that the autism diagnostic procedure can be decomposed into a set of screenings items where each one is able to quantitatively analyze certain behaviors of the children with machine learning methods. With the help of machine learning based techniques such as speech recognition, person detection/re-identification, face detection, facial pose estimation etc., we hope to reduce certain intermediate human-in-loop steps in the diagnosis procedure and generate machine assessment/predictions.

Identifying autism presents a challenging problem in the sense that certain signs of autism can be subtle and the discovery of these signs requires high-level understanding of human behaviors. For human, correct diagnosis generally requires long term professional training and considerable clinical expertise, and involves joint assessment of interactive behaviors from multiple sources such as conversation and activities. For machines, it is likely that the information from any single modality (e.g., speech signal or video) alone may not be adequate to support correct assessment and identification of autism. As a result, we propose to address this problem via multimodal behavior analysis, where we expect that the combination of different modalities will jointly embed richer information to benefit our task. We also hope that the large scale of data, together with data-driven methods, can increase the reliability of autism identification. To this end, we establish a large scale multimodal autism behavior analysis dataset which contains designed diagnostic procedures, interactive behaviors of children that are comprehensively recorded under multiple modalities and camera views, as well as human expert diagnostic scores used for dataset labels. Specifically, the diagnostic procedures are designed to follow three widely used clinical assessment protocols from the Autism Diagnostic Observation Schedule-Generic: 1. response to name, 2. separation and reunion, and 3. response to non-social sound stimulation. On top of this dataset, we further propose a series of machine learning pipelines that quantitatively analyze the behaviors of the examinees in each sub-task and give the corresponding assessment scores. In the experiment, we show that the proposed framework not only produces assessment results that highly correlates with human experts, but also gives autism identification predictions with promising accuracy.

3.2 Problem Setting

Although some previous methods have demonstrated promising performances on early screening, there are still many real constrains in their experiment setting. In particular, the most significant constraint in both experiments is the limitation of children's action. In the visual attention analysis experiment, for example, the test subjects are required to sit still in front of the screen for eye movement data collecting. It is worth noting that more natural interactions are expected to lead to more spontaneous behaviors and more accurate diagnosis. As a result, we seek to design an environment which gives children enough

spaces to interact naturally with multiple participants in this section.

One of the key symptoms of ASD patients is their impaired interpersonal communication ability[85]. The main target of this work will be focused on analyzing hevavior signals from multiple cameras. We seek to decompose the clinical diagnostic procedure into a set of activity items where one is able to quantitatively analyze certain behaviors related to ASD.

3.3 Multimodal Data Collection

To collect large-scale multimodal behavior data from children, we established an ASD diagnosis and behavior recording environment.

The general layout of our multimodal behavior recording system is shown in Fig. 3.1. The system consists of 8 HD cameras, 1 Kinect sensors and a 6-channel wireless audio capture system. The sets of sensors are expected to comprehensively cover various modalities of the signals from the child. The lab environment is shown in Fig. 3.2.

Both HD cameras and Kinect sensors provide additional visual, depth and skeleton cues to further help analyzing the activities, attentions, as well as responsiveness of a child. On the other hand, the audio capture system records speech dialogs and meaningful sounds which can be used to better interpret the activity contexts and provide segmental information for parsing the activities.

The lab setting and dataset collection took almost one year, with 112 children enrolled in the data collection. In these children, 54 have ASD, 28 are with developmental delay, and 30 are typically developed. The age is between 15months to 36 months old, average 24 months. The boy to girl ratio is 5:1.

Each child attended three different sub-tasks: separation and reunion procedure, response to name(response to social sound stimulation), and response to non-social sound stimulation. The clinical scores for each sub-task were evaluated by human experts. The target of establishing this dataset is to see whether artificial intelligent systems can approximate human diagnosis via automatically analyzing the recorded multimodal children behaviors.

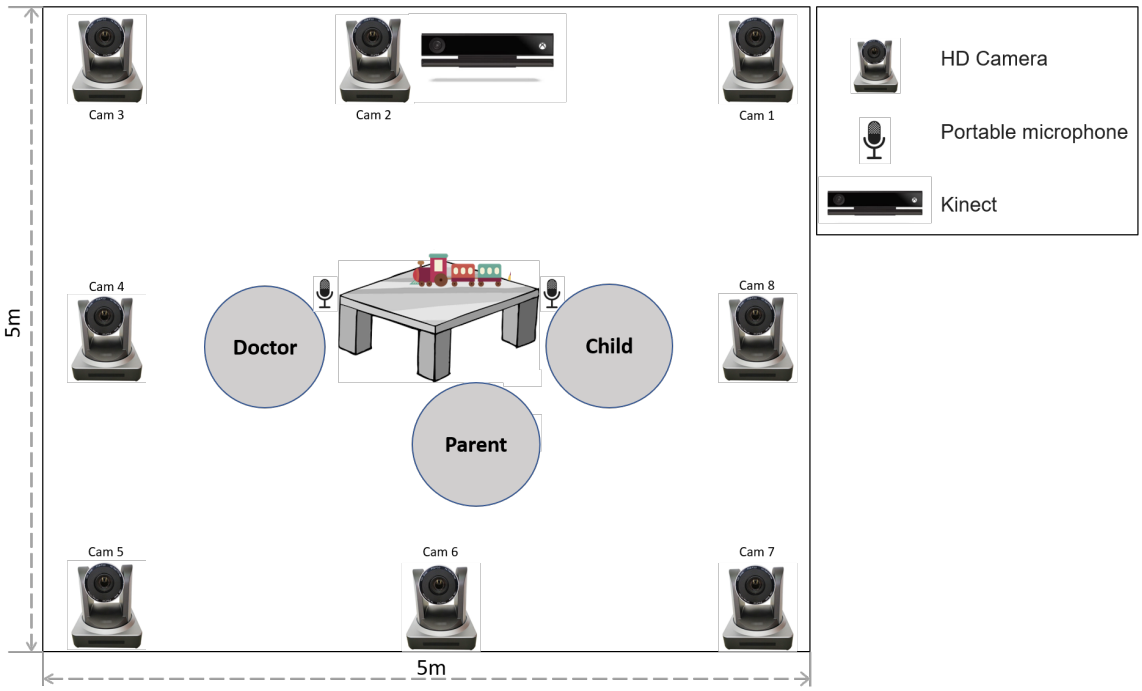

Figure 3.1: Illustration of the layout of the multimodal behavior recording system.

Figure 3.2: Data collection lab.

3.3.1 Response to name

Studies showed that individuals with ASD respond to their Names differently from typical developing (TD, which means non-ASD) ones, and the decreased tendency to name response has been found across studies [86]. In addition, such criteria is widely adopted in early screening and diagnostic assessments to identify early signs of autism. In particular, a clinician calls the name of a child, and scores based on whether and how quick a clear response can be observed. This motivates us to reproduce similar evaluation process with vision and learning techniques in order to reduce the required human interaction

and expertise. As illustrated in Fig. 3.3, the videos were collected in lab-controlled sessions with cleaner background and HD cameras. We hope that the children can react in a more natural way. We then propose a machine learning based multimodal responsiveness assessment framework. Our hope is to provide an non-subjective framework with minimized human interaction and expertise. While currently designed framework largely follows clinical tradition by looking into response time and duration, the proposed system is able to support future extended frameworks with abundant additional features, such as head pose, motion and pose-wise duration. The dataset collection procedure is ADOS-inspired, and follows similar protocol as the "response to name" probes in the ADOS assessment. In the experiment, each child is given a simple test where the child sits at a table playing with the accompanying adults or with toys. A doctor sits behind the child and calls his or her name in a natural tone as shown in Fig. 3.3 and Fig. 4.2. If the child turns around showing eye contact with the doctor, the test is completed. Otherwise, the doctor will repeat the call for additional 2 times. The whole process is recorded by a camera placed approximately behind the doctor (above the shoulder). The camera is expected to capture the eye contact from the child once he or she responds to the name calling.

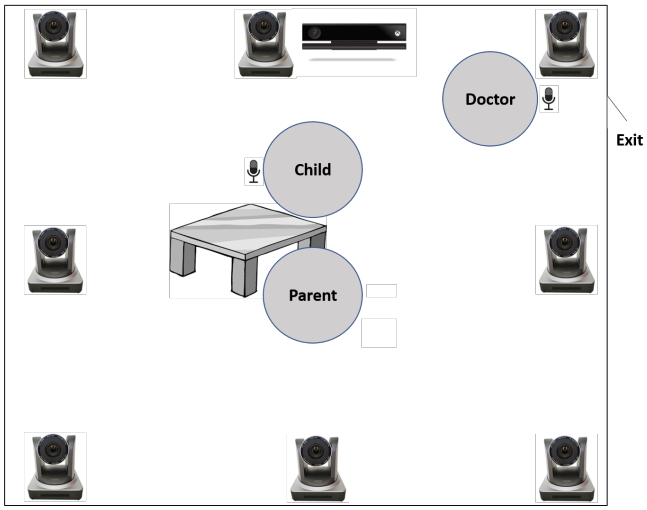

Figure 3.3: Lab layout of response to name task.

3.3.2 Separation and reunion

The second diagnose cue is attachment behavior. Based on Bowlby's theory of Attachment [87], separation from a primary caregiver is a natural cue to danger, leading to activation of the attachment system. Hofer

Figure 3.4: Perspective from Camera 1 (Left) and Camera 5 (right) in response to name task

also found that there may be multiple regulatory processes operative within the parent-child interaction, that may become temporarily dysregulated in separation status [88].

We create an environment (Fig. 3.5) in which the attachment behavior of the child can be observed via a separation and reunion procedure in a semi-structured setting. To this end, the data-collection is divided into two main steps as shown in Fig. 3.6:

Separation: In this step, the test subject sits with one parent and the doctor. The parent is asked to play with the child first, and then to leave the lab in the front of the child.

Response Observation: In this step, we analyze the response of the child by observing whether the child tends to follow the parent's trajectory or not.

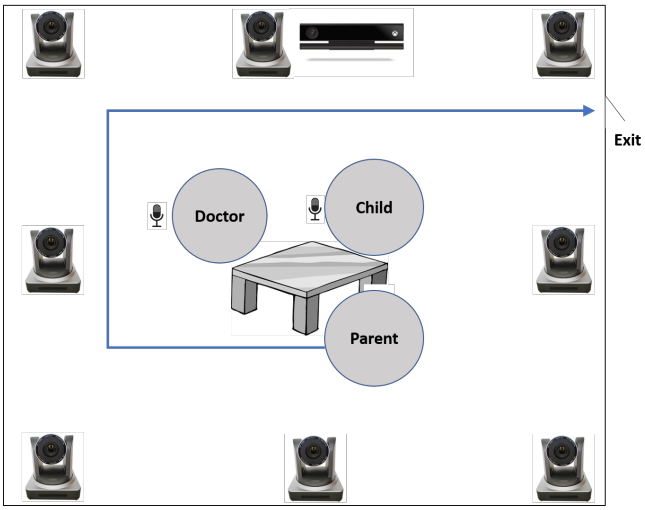

Figure 3.5: Lab layout of separation and reunion task.

In this test, we assume that a typically developed child tends to have some pro-social response in separation sub-task. The response of child during separation sub-task is classified into two classes which

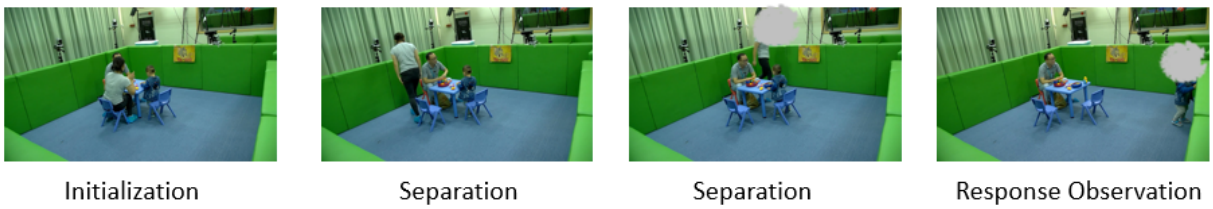

Figure 3.6: Illustration of separation and response observation.

are some pro-social response and little or no pro-social response. A some pro-social response of a child is defined as: the child orients towards parent, changes in behavior, looks around for the parent and tends to follow the parent. A little or no pro-social response is defined as: the child may or may not orient towards the parent but does not move towards the door. And the child continues with the activity without noticeable changes in behavior [89].

3.3.3 Response to non-social sound stimulation analysis

The third diagnose cue is response to non-social sound stimulation analysis. In this test, we assume that a child tends to point to or look at the objects which he has an interest in. A toy which can beep is hanged on the house beam (as shown in Fig. 3.7), and the sound can be remotely controlled by the doctor. During the experiment, the doctor use this non-social sound to stimulate the child. A positive response of the child should be pointing the toy with the hand as shown in Fig. 3.8. This response to non-social sound stimulation sub-task and response to name sub-task evaluate the reactions toward the source of non-social and social sound stimulation respectively.

In data collection: the child sitting next to a table concentrates on playing with toys in the beginning and an object was placed very close to a camera on one side as a non social sound stimulus before the test as shown in Fig 3.8. Considering the relatively far distance, we assume that any child's response to the sound stimulation is equivalent to response to the camera. When the sound stimulus starts, the typical developed child is expected to turn his head and look at the object. The clinician will also vocally instruct the child to point to the object if the child does not point to the object spontaneously. If the child does not follow the speaking instruction, clinician will then himself point to the object to guide the child and check whether the child will point to the object. The index-finger pointing to objects is an important metric since

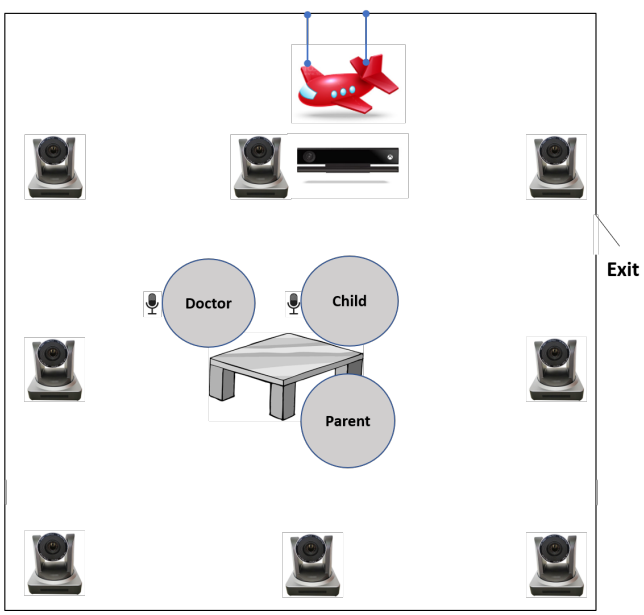

Figure 3.7: Lab layout of non-social sound stimulation sub-task.

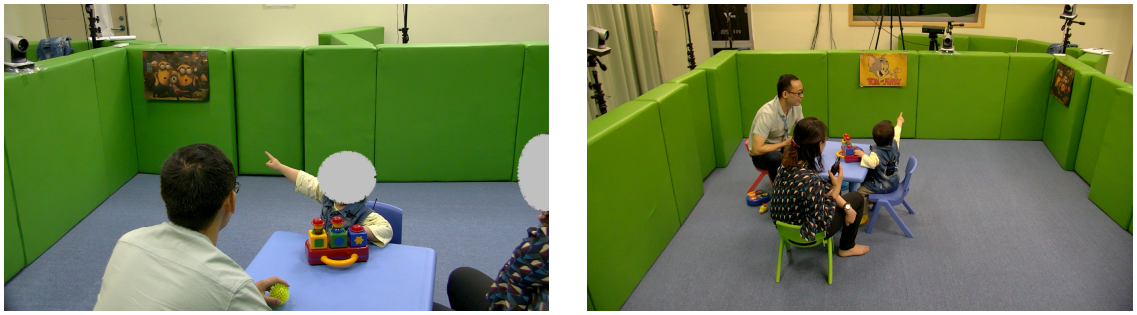

Figure 3.8: A positive response from the prospective of Camera 4 (Left) and Camera 6 (right) in non-social sound stimulation analysis sub-task. At first the child is playing toys on the table, and then a helicopter toy above the camera beeps. The child hears the sound, turns his head and look at the toy. Following words instruction, the child finally points to the object.

studies have examined that autistic children are less likely to use their index finger to point([90],[91]).

In Fig. 3.9, the child shows a negative response to the non-social stimulation. We can see that after clinician delivering instruction, the child is still playing the toys and showing no response to the clinician.

Figure 3.9: An illustration of negative response.

3.3.4 Collection procedure

The three sub-tasks are performed under the same data collection environment in the same trial. For each child, all tests are conducted in sequence, and minutes of unstructured plays are also introduced between different tests. The proposed dataset contains sessions with dynamic and challenging scenarios. Our expectation is to offer the children a more natural environment so that their reaction may better reveal their psychological status. As a result we do not impose too much control on the input video.

3.3.5 Clinical scoring

We ask a psychological clinician to evaluate each recorded video from each sub-task with a subjective score within the range of $[0, 1, 2]$. In general, in each sub-task a clear and fast response gives the score of 0, a partial response gives 1, while no response gives 2. The performance of each child in each sub-task is associated with one subjective score. It should be noted that although having considerable correlation, the above assessment is an evaluation of the responsiveness to different sub-tasks, not an evaluation of autism risk. For every child, he or she is labeled by three subjective scores from sub-tasks and one ASD diagnose score (1 or 0) after a complete set of diagnostic procedures.

3.4 Data Processing

In this section, we summarize the main data processing steps which our behavior analysis methods will be based on.

3.4.1 Automatic name calling detection

A core sub-task in "response to name" experiments is to locate the time stamps of name callings such that response latency can be measured. In [53], name calling is annotated manually and incorporates unnecessary human interaction. We therefore propose an automatic name calling detection system based on automatic speech recognition (ASR). In particular, we designed an ASR system based on Kaldi [41], a toolbox widely used for speech recognition. The acoustic model is trained from 1000 hours of Mandarin telephone conversations with the chain model setup in [41]. In our experiment, the name of each child is

first registered by our ASR system. The registered name is then matched with the ASR recognized speech signals throughout the video to locate the name callings.

3.4.2 Face detection and alignment

Another important sub-task is to localize the child's head since it is the major body part which conveys actions of response and eye contact. Response and eye contact are often characterized by facing towards the interacting person. Such assumption becomes particularly valid when children are suddenly called from behind while being highly focused. As a result, face detection presents a good method in signaling children's response when being called. In this work, we use the DLib [92] implementation of the face alignment methods proposed by Kazemi et al. [93] to simultaneously detect and align the faces. Besides detecting faces, the algorithm returns 68 landmarkers which later will be used to compute the head pose. An example of the detected face landmarkers is shown in Fig. 3.10.

3.4.3 Face verification

Since we allow accompanying adults, sometimes they may also respond to name calling with their faces detected. As a result, multiple faces from both adult and child can simultaneously appear in the same frame, and face verification is needed to distinguish the desired children's faces. Again we register each child's face in the system and verify every detected face based on the method proposed in [94]. We also formulate the verification problem as a structured sequence prediction problem with temporal information incorporated. This helps to stabilize and improve the verification performance under certain situations.

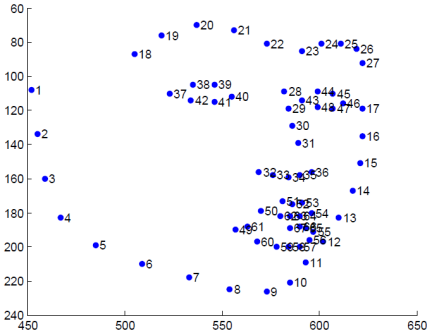

Figure 3.10: The facemarker demonstration

3.4.4 Pose estimation and tracking

To analyze the body behaviors, we use AlphaPose [65, 66, 67], a state-of-the-art pose estimation and tracking library to predict and track the 2D skeletons of the test subjects in our videos. AlphaPose follows a two-step pose estimation framework which uses a VGG-based SSD-512 detector [95] to detect humans, followed by a spatial transformer network (STN) and a single person pose estimator [96] to adjust the box localization and output human poses.

Figure 3.11: Examples of detected and tracked poses using AlphaPose.

The base method of AlphaPose described above is frame-wise independent. The library also provides a tracking feature based on pose flow [67]. We adopt this tracking feature for our work, which groups individually detected poses into temporally consistent tubes. Fig. 3.11 shows some examples of the detected and tracked poses obtained by AlphaPose on our videos.

Pose estimation and tracking is the most important part in our pipeline. It serves as the purpose for the joint localization, tracking and pose abstraction of the test subject. This module thus provides the mid-level perception results for two sub-tasks described in next chapter - the "separation and reunion" test and the "response to non-social sound stimulation" test.

3.4.5 Person re-identification (re-id)

Pose estimation and tracking can only detect and temporally group the estimated poses into tubes, but do not have the capability to associate them with identities. Since multiple people including the child can be present at the same time, it is important to use person re-id to associate such identity by matching the tubes to registered person images from the beginning.

To this end, we proposed a person re-id framework with a spatial-channel co-occurrence model, in conjunction with deep networks to form an end-to-end learnable re-id framework. We motivate the proposed method by showing an example below. A conventional way to compute the similarity between two images is to take the mean pooling of each channel in their feature maps, followed by taking a cosine similarity between the pooled feature vectors. As shown in 3.12, a brittle aspect of this design is that it requires channels to be well aligned. However, when channel are not well aligned, such similarity measurement does not capture the high similarity in the patterns.

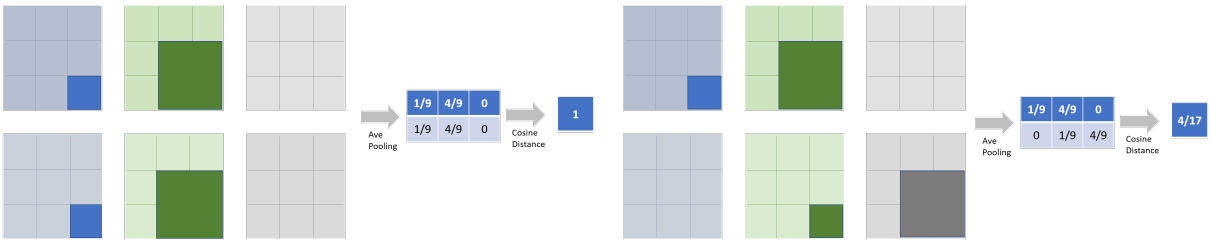

Figure 3.12: An example of conventional computation of the similarity between two feature maps. Left: channels are well aligned. Right: channels are not well aligned.

As shown in Fig. 3.13, the key idea of the proposed co-occurrence model is to consider the correlation across different feature channels. This is done by computing the inner products between one feature map channel and different channels from another feature map, and take them all into account. In this way, the proposed similarity measurement is invariant to the shift of channels.

In practice, we embed the above design as a cross-channel correlation layer at the end of a backbone to incorporate the above inductive bias, as illustrated in Fig. 3.14. This layer takes the pairwise inner products between channels as the activation, forming a vectorized input taken by subsequent fully connected (FC) layers before the final prediction.

It should also be mentioned that our proposed model not only considers the cross-channel informa-

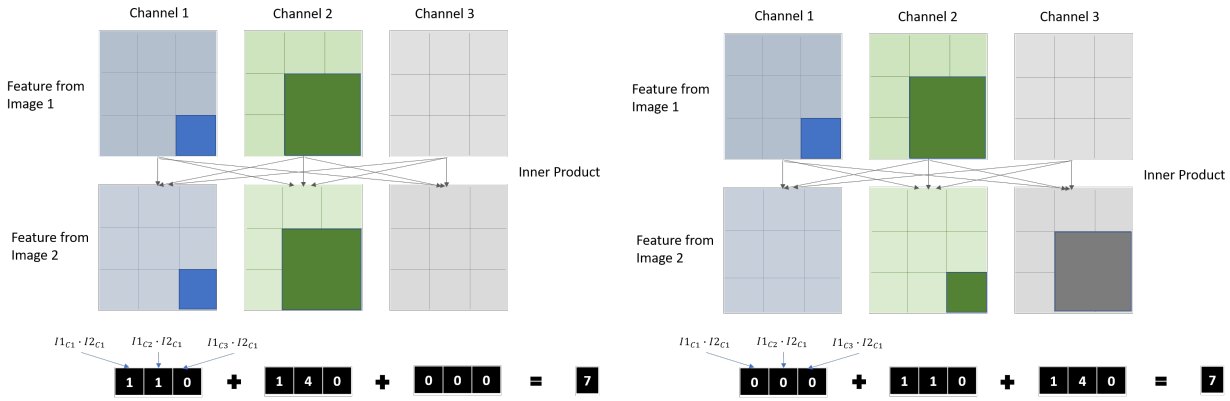

Figure 3.13: An example of the proposed similarity measurement between two feature maps. Left: channels are well aligned. Right: channels are not well aligned.

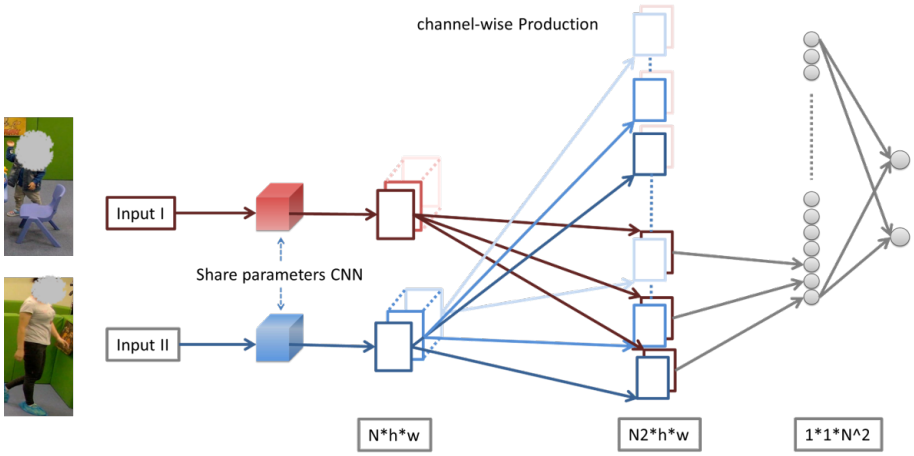

Figure 3.14: Illustration of the proposed person re-id architecture.

tion, but also takes into consideration the spatial domain cross-pixel information. The above cross-channel correlation layer design is thus similarly applied to the spatial dimension, where cross-pixel feature inner products are computed similar to the cross-channel scenario. Specifically, the spatial dimension is stretched into an $H \times W$ vector. In this case, the channel dimension is also a vector and the inner product is taken between the channel vectors. These two branches are finally fused at the loss level during training via a weighted sum of the individual softmax losses on each branch. This design allows the proposed method to have more tolerance to large pose variations, in contrast to conventional methods that assume relatively more rigid object appearance.

We adopt ResNet-50 as the backbone following existing literature, and pretrain the network on Market-1501 [97] which is a large-scale person re-id dataset. Similar to pose estimation and tracking, this transfer

learning setting allows us to leverage existing large-scale data to benefit our task and overcome the limit of data size. The ResNet-50 backbone is also pretrained on ImageNet-1k. The pretraining aims for general object recognition but the learned representations still reflect object's elementary features. It thus also helps the performance and robustness of our person re-id module.

Chapter 4

Identifying Children ASD via Multimodal Behavior Signal Analysis

In this chapter, we describe the proposed machine-based evaluation pipelines for multimodal behavior analysis. The pipeline contains three sub-tasks described as follows:

4.1 Sub-task 1: Response to Name

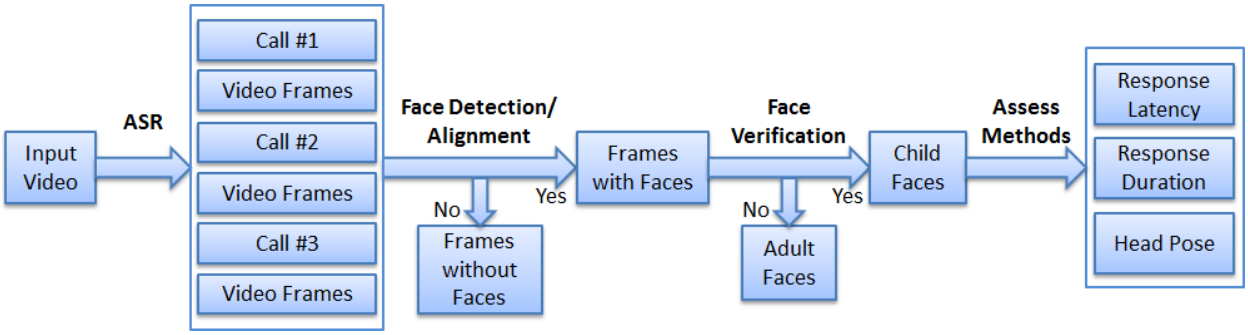

Figure 4.1: The proposed multimodal machine learning framework towards"response to name"

An overview of the proposed multimodal "response to name" assessment framework is shown in Fig. 4.1. In this work, both speech processing and vision based methods are incorporated to minimize the human annotations in assessing the responsiveness. In particular, we simultaneously consider response speed, response duration as well as head pose information to jointly determine a predicted score. We hope such method can help to discover atypical response patterns effectively.

Figure 4.2: A typical positive response to name illustration

4.1.1 "Response to name" assessment

With the detected child faces, we are able to propose a rule-based "response to name" assessment framework. Similar to conventional clinical diagnosis, our framework is based on the following two basic assumptions:

Assumption 1: A clear response should happen with relatively small latency upon calling.

Assumption 2: A clear response should last for a certain length of duration.

The response latency in the first assumption can be naturally modeled as:

$$latency = T_f - T_{c1}, \tag{4.1}$$

where T_f and T_{c1} indicate the time stamp of the first detected face and the beginning of the first call, respectively. The response duration in the second assumption can also be modeled as:

$$duration = N_f/\text{frame rate}, \tag{4.2}$$

where N_f represents the total number of frames containing detected child faces.

We observe that head pose also presents an important source of response information. Not only is the pose correlated with the clearness of response, but also it can provide head motion information that will benefit future assessment methods. Unfortunately, the DLib face alignment algorithm does not provide such information. We therefore propose a light and effective real-time pose estimation based on

the detected landmarkers. We first propose a robust feature to effectively encode the pose information. Suppose there are n landmarkers whose coordinate in the frame is represented as (x_i, y_i), the head pose feature can be computed as:

$$\begin{aligned} f = [&x'_1 - x'_2, ..., x'_1 - x'_n, x'_2 - x'_3, ..., x'_{n-1} - x'_n \\ &y'_1 - y'_2, ..., y'_1 - y'_n, y'_2 - y'_3, ..., y'_{n-1} - y'_n] \end{aligned}, \tag{4.3}$$

where x_i and y_i are the normalized relative coordinates of the ith marker with respect to the left and top most land markers:

$$\begin{aligned} x'_i &= \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}} \\ y'_i &= \frac{y_i - \min\{y_i\}}{\max\{y_i\} - \min\{y_i\}} \end{aligned}. \tag{4.4}$$

Since the above feature looks into the differences of all non-repeated pairwise landmarker combinations, the feature dimension can be high. This significantly adds computational costs to our head pose algorithm. We note that a large portion of the dimensions in fact contains redundant and non-informative information. As a result we apply PCA to the extracted features and maintain the top 20 dimensions with the highest energy. Given a training set with labeled head pose angles and any testing set, we perform face detection/alignment on both sets and extract the above head pose features. We then regress the head pose on the test set by referring to the top K nearest training samples with majority pose voting.

Upon obtaining the head pose information we incorporate it in the duration estimation by weighting each frame with the following biased Gaussian kernels:

$$duration = \sum_{i \in \mathbf{F}} k(\theta_{x,i}, \theta_{y,i}) / \text{frame rate}, \tag{4.5}$$

$$k(\theta_x, \theta_y) = \exp(\frac{\theta_x^2}{2\sigma_x^2}) \exp(\frac{\theta_y^2}{2\sigma_y^2}), \tag{4.6}$$

where $\theta_{x,i}$ and $\theta_{y,i}$ are the horizontal and vertical head pose angles of the ith detected child face. In our experiment, we set σ_x and σ_y respectively to 45 and 60.

4.1.2 Score prediction

We perform grid search on both *latency* and *duration*, and optimize the classification and regression trees to predict the responsiveness score. We will introduce the accuracy of prediction clinical score in

experimental result section.

4.2 Sub-task 2: Separation and Reunion

To analyze whether a child demonstrates pro-social response, we extract certain continuous features to measure the duration and strength of a child's response. This differs from 0/1 scores as they provide more fine-grained information to measure the pro-social level. Specifically, we consider several quantitative indicators as shown in Fig. 4.3: 1) distance between the child and the parent, 2) distance between the child/parent and the door, and 3) the child's length of stay at the door.

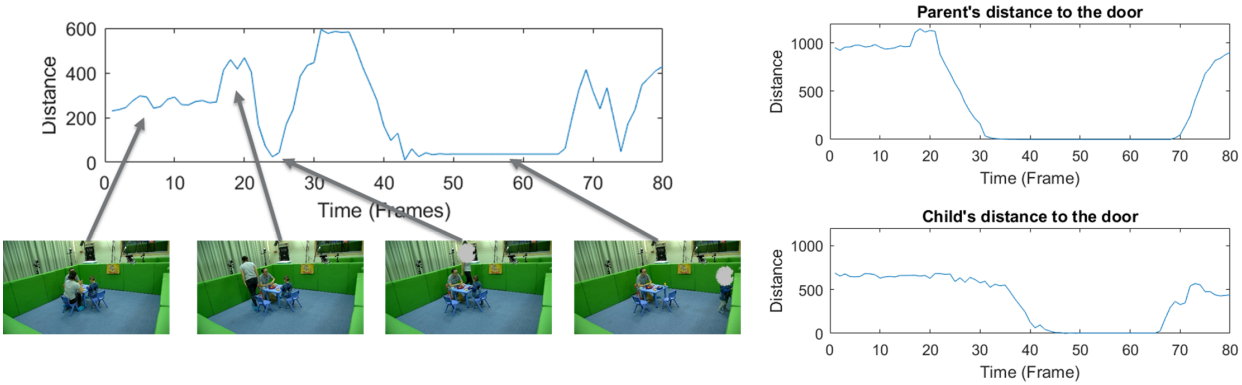

Figure 4.3: Quantitative indicators to measure the pro-social level of a child's response in the separation and reunion task.

This trajectory analysis problem can be decomposed into pedestrian detection and re-identification part as shown in Fig. 4.4. In the pedestrian detection step, we follow [98] where we use a region proposal network (RPN) to generate boxes and corresponding features, followed by a boosted forest to classify whether the proposed regions are people or not. We then use the person re-id framework proposed in the previous chapter to identify whether the detected person is a child or parent.

4.3 Sub-task 3: Response to Non-social Sound Stimulation

An important part of ASD diagnosis is the response to non-social sound stimulation test which analyzes the test subject's ability in reaction and expression. When playing with toys, non-ASD subjects are likely to get surprised and turn around if a sudden non-social stimuli sound occur, while ASD subjects, on the contrary, may continue their original work or pay few attention to the source of the sudden non-social

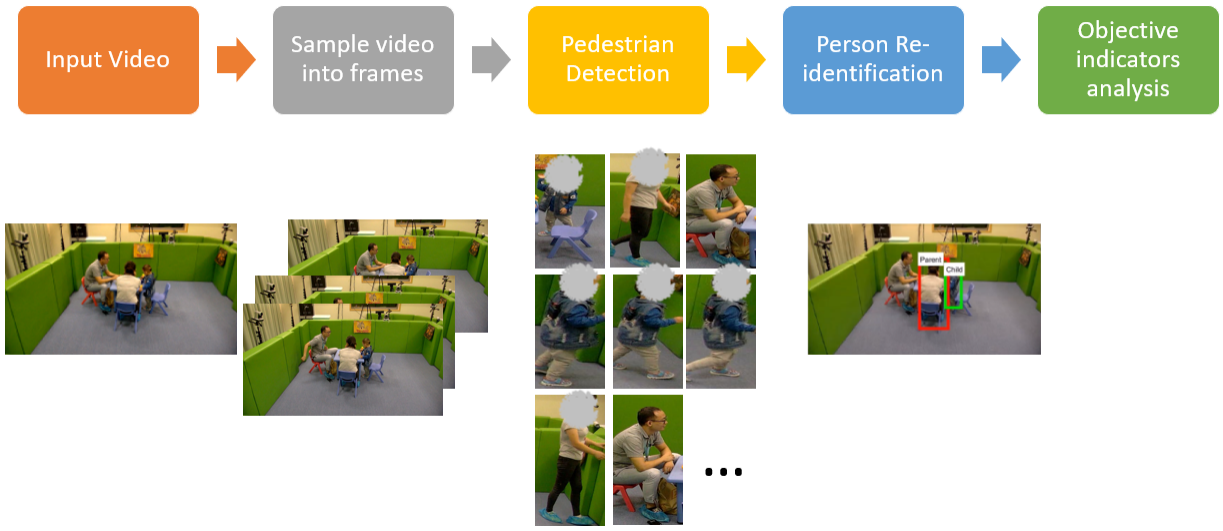

Figure 4.4: Illustration of the separation and reunion score prediction Framework.

sound stimulation. In conventional diagnostic procedures, professional clinicians are needed evaluate children's performance and score them based on their observations.

For machine based evaluation, we need to first localize the test subject. This is done by the pose estimation, tracking and person re-id pipeline as mentioned in the previous section. Person re-id associates the identities to the tracked pose tubes, which allows us to localize the subject by finding the corresponding tube. We then identify whether the subject has response by modeling the patterns of extracted poses using deep neural networks.

4.3.1 Problem formulation

We formulate the recognition of children response as a supervised event detection problem. In our experiment, each video begins from the non-social sound stimulation and ends by the end of the sub-task. And we assume that the videos are temporally segmented by labels that indicate whether the subject is responding or not. An important goal of this sub-task is to design a recognition system that can learn to detect the test subject's response action at a frame level.

We achieve the above goal by leveraging both spatial and temporal information jointly. To this end, each video is divided into overlapping short clips, as shown in Fig. 4.5. Each clip contains a short sequence of frames which form a window centered at the key frame. The label (subject responding or not) of the

clip is determined by the label of the center key frame.

It is worth noting that another major goal here is to output a machine predicted score of responsiveness given the whole video. Thus an ensemble of the predictive results from each clip is necessary. A video containing a responsive test subject dose not mean that all clips from this video are positive. Instead, clips from a positive video may correspond to both positive and negative actions. It often happens that the subject may look around and do not show response until the k-th clip of the video. Such untrimmed setting makes the problem more challenging.

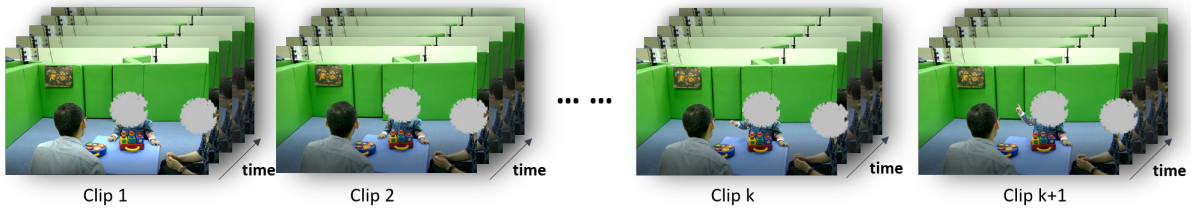

Figure 4.5: The demonstration of relationship between video label and clip label.

4.3.2 Representation with binarized skeleton map

We propose a simple but effective method to encode the subject's response. The result returned by AlphaPose contains a sequence of tracked human body keypoints, and conventional skeleton-based action recognition methods often resort to the modeling of inter-keypoint graph relation [99]. This way of modeling is effective given that graph relation between keypoints is the most direct abstraction of human actions. But such process is also baked in a lot of hand-crafted ingredients because the choice of graph relation is usually based on intuitions and heuristics.

We take an alternative approach where instead of trying to model the graph relation of keypoints, we directly generate binarized skeleton maps as the image input to a deep network. Examples of the binarized skeleton maps from both responding subjects and non-responding subjects are shown in Fig. 4.6. There are several advantages with this form of representation: 1) The binarized skeleton is an abstracted representation of the response that filters out irrelevant information. 2) The learning pipeline becomes simple and convenient, allowing us to leverage existing mature CNN architectures for this task. 3) The design partly avoids the hand-crafted nature of existing methods as mentioned above.

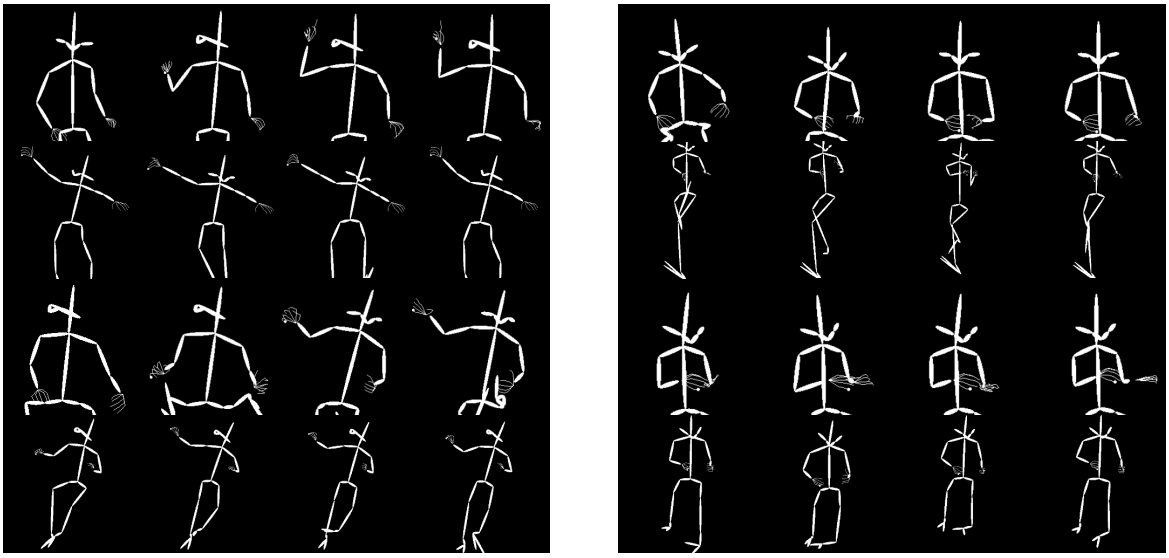

Figure 4.6: Visualization of binarized skeletons. Left: skeletons of the responding subjects. Right: skeletons of the non-responding subjects.

It should be mentioned that using binarized skeleton maps can be also interpreted as a transfer learning pipeline that leverages the rich pretraining data for pose estimation. These data are highly relevant to our task, and helps to overcome the limitation of data size in our problem.

4.3.3 Learning and evaluation pipeline

Once obtaining the binarized skeleton maps, we can treat them like an "MNIST-like" classification dataset. Instead of taking in single images, we form overlapping clips as mentioned in Sec. 4.3.1, and use a convolutional neural network with frame stacking to predict the clip labels. An illustration of the learning pipeline is shown in Fig. 4.7. Specifically, we use a 5-layer LeNet as the shared network to encode the features before pooling. For frame stacking, we consider a max pooling operation to fuse the embeddings across different frames in each clip.

Since our goal is to output machine evaluation of responsiveness at a test subject level, ensemble of the independent clip level predictions is needed. For a complete video, simply taking the maximum response score from the set of clips may lead to overestimate of the responsiveness due to false positives. We thus take both the duration and the strength of response into consideration, by defining the subject-level score as the sum of clip predictive scores which are higher than a threshold T.

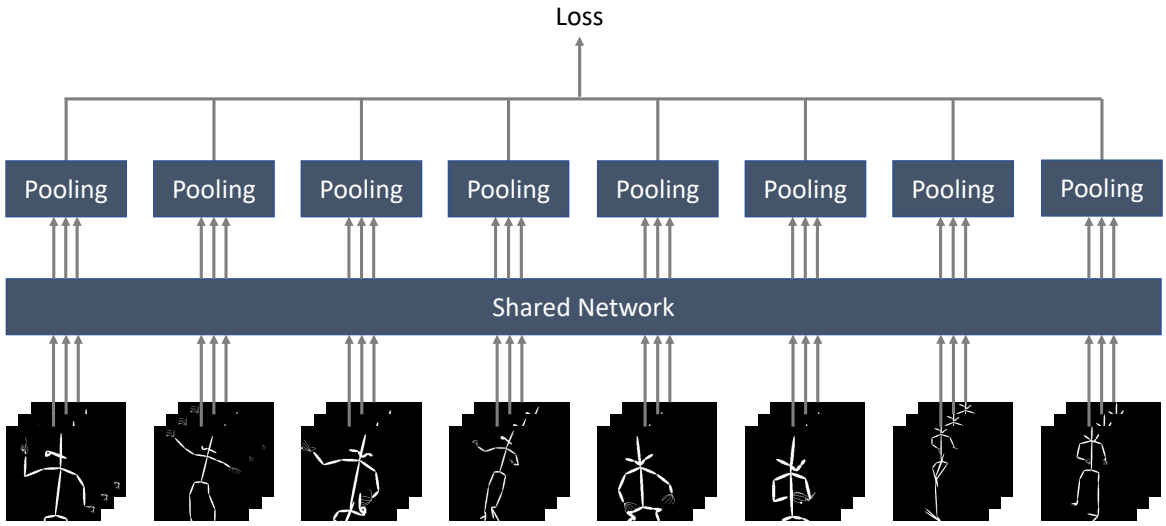

Figure 4.7: Recognizing the clip labels with a convolutional neural network.

Due to the limitation in data size, we adopt a "leave-one-out" cross validation strategy to split the videos and evaluate each video (subject) in an iterative manner.

4.4 Experiments

In this section, we discuss the comprehensive experiments and analyses corresponding to the sub-tasks defined in the previous sections. Overall, our experiments are structured in a way to study the following three aspects: 1) The correlation between machine evaluation and human evaluation on each sub-task. 2) The accuracy using the machine evaluation or human evaluation as individual ASD predictors. 3) The effect of model ensemble in ASD prediction.

The first aspect intends to examine the reliability of our perception and machine evaluation frameworks, by comparing machine evaluation result against doctor labeled response scores. For each sub-task, the clinical score is divided into 2 cases: 0 for positive response, and 1 for negative response. And we use confusion matrix to examine the correlation between machine evaluation and human evaluation.

The second aspect intends to study the role of each sub-task in the final prediction of ASD. Our hope is to understand the efficacy of the sub-task designs, which echoes our effort to establish a machine evaluation protocol similar to existing ones in clinical practices.

The third aspect intends to understand whether the fusion of different tasks, or different sources of

evaluation, can lead to improved ASD diagnosis performance. However, it should be mentioned that proposing a machine based ASD diagnoses framework is not the sole purpose of this work. Another important goal is to understand the role of machine generated scores in assisting professional diagnoses. Thus besides the ensemble of machine evaluation from different sub-tasks, we also consider scenarios with a hybrid source of scores from both machine and clinicians. Finally, we incorporate the Modified Checklist for Autism in Toddlers (M-CHAT) to further assist the assessment of ASD risk. M-CHAT is a psychological questionnaire which aims to identify 20 behavioral characteristics of the autism spectrum and asks if the child has experienced any of them. It makes a good complement to our lab-controlled tasks. We jointly consider all the sub-tasks with M-CHAT scores together, with an illustration shown in Fig. 4.8.

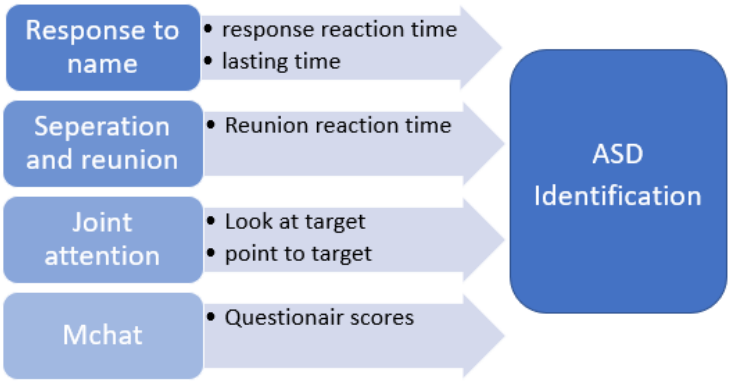

Figure 4.8: The general procedure to determine whether a child is ASD patient. Four different tasks with multiple features are considered.

Note that for any experiment involving the learning of a classifier, we follow the same "leave-one-out" strategy described in Section 4.3.3.

4.4.1 Response to name test

In this task, a positive response should have relatively short response latency and long duration of eye contacting. In other words, a typically developing test subject should respond to his name quickly and face the name caller for some time. We thus use response latency and front facing duration as our features, and apply the rule-based decision tree separately on each feature. Table 4.1 shows the confusion matrix of the machine evaluation score comparing against the doctor labeled response score. We can see that the

proposed method achieves an accuracy of 93.7% using the doctor labeled response score as ground truth.

Table 4.1: The confusion matrix of machine evaluation compared to doctor labeled response score in the response to name test.

		Prediction Score	
		Positive	Negative
Doctor Labeled Response Score	Positive	138	9
	Negative	5	70

Table 4.2: The accuracy, sensitivity and specificity of machine evaluated response score using doctor labeled response score as ground truth in the response to name test.

Accuracy	Sensitivity	Specificity
0.937	0.939	0.933

We also evaluate the performance of machine evaluation score in direct ASD prediction. In our test protocol, the parent is first asked to participate name calling, followed by the doctor. Thus there are two response to name tests performed on each test subject. Our results in Table 4.3 and 4.4 indicate that the performances based on the machine evaluation scores from parent and doctor name callings are 70.3% and 63.1%, respectively. Table 4.5 further shows the ASD prediction result obtained by concatenating the scores of both name calling and using a random forest classifier, achieving 73.9% accuracy. The improved performance indicates that the machine scores from both name callings can be complementary. And the accuracy, sensitivity and specificity of ASD prediction in response to name task is shown in Table 4.6.

Table 4.3: The confusion matrix of ASD prediction based on parent name calling.

		Prediction Score	
		Positive	Negative
ASD Label	Positive	37	21
	Negative	20	33

Table 4.4: The confusion matrix of ASD prediction based on doctor name calling.

		Prediction Score	
		Positive	Negative
ASD Label	Positive	41	17
	Negative	16	37

4.4.2 Separation and reunion test

In the separation and reunion task, a positive response means that the child will follow the parent immediately after noticing the leave of the parent. We extract the trajectory of both the parent and the child,

Table 4.5: The confusion matrix of ASD prediction based on the fusion of both parent calling and doctor calling with a random forest.

		Prediction Score	
		Positive	Negative
ASD Label	Positive	33	19
	Negative	11	42

Table 4.6: The accuracy, sensitivity and specificity of ASD prediction in response to name task.

	Accuracy	Sensitivity	Specificity
ASD prediction based on parent calling	0.631	0.638	0.623
ASD prediction based on doctor calling	0.703	0.707	0.698
ASD prediction with fusion of both sessions	0.714	0.635	0.792

and obtain the time delay of child's arrival to the door after parent leaving the room. The performance evaluation is based on their trajectory: if parent's and child's coordinates in the video coincide with each other or the child arriving at exit after parent leaves, the behavior is labeled as positive. As shown in Table 4.7, the accuracy of machine evaluation compared to response score is 87.39% as shown in Table 4.8. Using this score, the performance of ASD prediction is 56.8%, as shown in Table 4.9 and Table 4.10.

Table 4.7: The confusion matrix of machine evaluation compared to doctor labeled response score in the separation and reunion test.

		Prediction Score	
		Positive	Negative
Doctor Labeled Response Score	Positive	65	11
	Negative	3	32

Table 4.8: The accuracy, sensitivity and specificity of machine evaluated response score using doctor labeled response score as ground truth in the separation and reunion test.

Accuracy	Sensitivity	Specificity
0.873	0.855	0.914

4.4.3 Response to non-social sound stimulation test

In this task, we analyze the subject's reaction to non-social sound stimulation based on the pose estimation and tracking pipeline proposed in Section 4.3. Table 4.11 and Table 4.12 shows that machine evaluation achieves an accuracy of 81% when compared to doctor labeled response score. In addition, Table 4.13 and Table 4.14 shows that the performance of using the machine evaluation score in this task alone leads to 76.6% accuracy in ASD prediction.

Table 4.9: The confusion matrix of ASD prediction based on separation and reunion test.

		Prediction Score	
		Positive	Negative
ASD Label	Positive	23	35
	Negative	13	40

Table 4.10: The accuracy, sensitivity and specificity of ASD prediction in the separation and reunion test.

Accuracy	Sensitivity	Specificity
0.568	0.397	0.755

Table 4.11: The confusion matrix of machine evaluation compared to doctor labeled response score in the response to non-social sound stimulation test.

		Prediction Score	
		Positive	Negative
Doctor Labeled Response Score	Positive	51	6
	Negative	14	40

4.4.4 M-CHAT score

Besides machine-based evaluation, we also consider the Modified Checklist for Autism in Toddlers (M-CHAT) questionnaire as an additional input. M-CHAT is one of the most important ASD pre-screening methods that has been widely adopted. The questionnaire is to be filled by the parents and includes 23 yes/no questions.

In the experiment,the 23 yes/no result of one subject is marked as 0 and 1 and concatenated as one 23-dimensional feature. Similar with previous experimental setting, the data was divided into training and testing set with "leave-one-out" cross validation strategy and then we evaluate each subject in an iterative manner. With random forest classifier, the ASD prediction accuracy is 82%, and the confusion matrix is shown in Table 4.15.

Based on the work in [100], M-CHAT alone could predict 69% of the autism cases for toddlers aged 20 months and older, 36% for toddlers younger than 20 months. From our experimental result, we could find that machine learning based ASD prediction with M-CHAT shows promising results with the accuracy of 82%, sensitivity of 89.7% and specificity of 73.6%.

Table 4.12: The accuracy, sensitivity and specificity of machine evaluated response score using doctor labeled response score as ground truth in the response to non-social sound stimulation test.

Accuracy	Sensitivity	Specificity
0.81	0.895	0.741

Table 4.13: The confusion matrix of ASD prediction based on response to non-social sound stimulation test.

		Prediction Score	
		Positive	Negative
ASD Label	Positive	39	19
	Negative	7	46

Table 4.14: The accuracy, sensitivity and specificity of ASD prediction based on response to non-social sound stimulation test.

Accuracy	Sensitivity	Specificity
0.766	0.672	0.868

Table 4.15: The confusion matrix of ASD prediction based on M-CHAT.

		Prediction Score	
		Positive	Negative
ASD Label	Positive	52	6
	Negative	14	39

4.4.5 Model ensemble and M-CHAT score feature selection

We notice that the performance of ASD prediction based on individual task is not good enough. Clinical ASD diagnoses are typically conducted based on the doctor's subjective observations on multiple sub-tasks. In each sub-task, a child's behavior is evaluated as either high risk or low risk. However, each single high risk behavior alone cannot identify ASD reliably. For example, a non-ASD child sometimes may not have any response to the name callings, but an ASD child may have positive response to name callings because of different specific personalities. Therefore ASD children need to be identified by considering different behaviors together.

To this end, we evaluate the ensemble of three different sub-tasks, by concatenating their machine evaluation scores and learning a random forest classifier.

Table 4.17 shows the confusion matrix of ASD prediction with an accuracy of 81.1%.

Finally, we are interested in investigating how machine evaluation can assist the ASD diagnosis together with response score and what information in M-CHAT is useful for machine-based screening.

To this end, we treat the M-CHAT answer of each question as a feature and perform incremental

Table 4.16: The accuracy, sensitivity and specificity of ASD prediction based on M-CHAT.

	Accuracy	Sensitivity	Specificity
ASD Prediction	0.820	0.897	0.736

Table 4.17: The confusion matrix of ASD prediction based on combination of all 3 tasks.

		Prediction Score	
		Positive	Negative
ASD Label	Positive	47	11
	Negative	10	43

feature selection. We greedily add a feature into an existing set of selected features based on the prediction performance after adding it in. This ensemble framework leads to a 89.2% accuracy in ASD prediction. Our method outperforms the conventional early screening method by a significant margin. Table 4.18 shows the confusion matrix of the above ensembled model.

Table 4.18: The confusion matrix of ASD prediction based on combination of machine score and M-CHAT score

		Prediction Score	
		Positive	Negative
ASD Label	Positive	54	4
	Negative	8	45

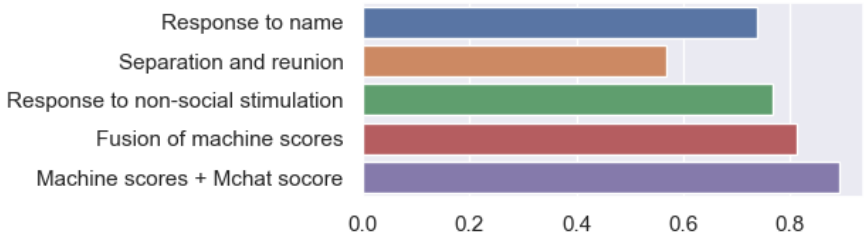

Figure 4.9: The comparison of different experimental settings.

In our model we found that adding 3 dimensions of the M-CHAT score already leads to the optimal performance (89.2%). The selected 3 dimensions correspond to the following questions:

- 11. Does your child ever seem oversensitive to noise? (e.g., plugging ears)
- 15. If you point at a toy across the room, does your child look at it?
- 19. Does your child try to attract your attention to his/her own activity?

We observe that the selected 3 questions are complementary to our designed tasks. An example is the last question "Does your child try to attract your attention to his/her own activity?" In our existing experiments, the proposed lab-control tests are focus on social sound stimulation (response to name),

non-social sound stimulation, and parent-child relationship characteristics (separation and reunion task). However these automatically selected 3 question set focus on test subject's proactive behavior. We observe the response of all test subjects by letting them receive external stimuli in a passive manner. Our task thus has not considered active behaviors of the test subject, which is exactly covered by this question. Such observation can motivate your future task design.

4.5 Discussions and Remarks

4.5.1 Experimental observations

Our experiments lead to some observations: 1) When comparing 3 different sub-task, we could find that the separation and reunion task performs worst, and the results in this sub-task are more random. We observe that for both ASD or TD children, if their attention is focused on the toy, they tend to not follow their parents. 2) The person who participates the test matters. For example, in the response to name task, the experiment conducted by the doctor leads to improved ASD identification accuracy.

4.5.2 Clinician guidance

our work can help the clinicians guide their tests and analyses in the following aspects: First, the quantitative evaluation of each single behavior gives doctor an objective indicator of the child's behavior. This result is not influenced by the subjective evaluation of different doctors. Second, automated machine evaluation helps doctor to reduce their efforts on recording fine-grained children behaviors. For example, in the response to name task, doctors often sits on the side of children when parents call the name. Thus it is more difficult for them to always observe the children expression and attention precisely. In this case, machine evaluation presents a great complementary that allows doctors to gain more information with less efforts. Third, the comparison of different tasks can identify which behavior has more significant differences between ASD and TD children. The doctor can thus pay more attention and design the paradigms following the guidance of machine scores.

4.5.3 Future paradigm designs

In our research, we only used 3 paradigms to achieve the current performance. Additional paradigms would allow the observation of more behavior patterns of a child, and would thus lead to improved accuracy. The design of new paradigms should be complementary to the existing ones. Based on the results of model ensemble and M-CHAT feature selection, we could find that current paradigms focus more on a child's reaction under different social and non-social stimulation. However, our task has not considered active behaviors of a test subject where more new paradigms can be designed.

4.5.4 Complementary relation between M-CHAT and machine evaluation

While machine based evaluation can output objective information in lab environments, the children's reactions from daily life observed by the parents are also very important. Without lab controls, children's behaviors at home with their parents are more spontaneous, and the parent's observation over a long period can provide more abundant information. This observation is reflected in M-CHAT, making it a good complementary to machine evaluation.

Chapter 5

Conclusions

In this work, we propose a non-intrusively multimodal behavioral signals capturing system for early screening of child ASD. The proposed system not only benefits human diagnosis by providing comprehensive data recording for observation purposes, but also serves as a starting basis for studies on data-driven child ASD screening. With multiple cameras covering the interaction area with different views, the system can effectively overcome occlusions and render diversified visual input towards more reliable behavior analysis. Using this system, we collected spontaneous multimodal behavior data from more than 110 children, and manually annotated these data in each sub-task by clinical doctors.

Our interaction protocol during data collection is motivated by the semi-structured assessment of ADOS. The protocols are designed to analyze ASD children's behaviors on social/non-social sound stimulation and parent-child relationship. With the collected data, we interpret the behavioral signals and propose a series of machine learning/deep learning based framework for child ASD identification. Although the proposed framework serves as a simple data-driven baseline, current experimental results show that such computer-based behavior analysis is promising. The machine predicted scores are highly consistent with clinical scores with more than 80% accuracy within each separate sub-task, and the system achieves an overall accuracy of 90.8% in leave-one-out ASD identification. We also use M-CHAT checklist as complementary input to cover different aspects of child behavior. We found that M-CHAT scores help to further boost our identification performance. This motivates us to design future experiment protocols based on items from the M-CHAT checklist.

Overall, our study provides promising findings on the possibility of computer-aided ASD early screening. There are many practical challenges for ASD early screening under current clinical practice. Despite its early onset, ASD is usually diagnosed several years later, mainly based on interviews with parents and clinical behavioral observations. Our proposed method takes less than 10 minutes for each child, and outperform traditional early screening methods with a significant margin.

Bibliography

[1] T. A. C. in Action, Autism prevalence is now 1 in 44, signifying the eighth increase in prevalence rates reported by the cdc since 2000.
URL https://www.prnewswire.com/news-releases/autism-prevalence-is-now-1-in-44-signifying-the-eigh
html 1

[2] L. Kanner, et al., Autistic disturbances of affective contact, Nervous child 2 (3) (1943) 217–250. 1, 4

[3] U. Frith, Autism: Explaining the enigma, Blackwell Publishing, 2003. 1

[4] K. Gotham, A. Pickles, C. Lord, Standardizing ados scores for a measure of severity in autism spectrum disorders, Journal of autism and developmental disorders 39 (5) (2009) 693–705. 1

[5] R. Zeligs, Glimpses Into Child Life: The Twelve-year-old at Home and School, W. Morrow, 1942. 2

[6] E. Bleuler, The theory of schizophrenic negativism, no. 11, Journal of nervous and mental disease publishing Company, 1912. 4

[7] H. Asperger, Die „autistischen psychopathen" im kindesalter, Archiv für psychiatrie und nervenkrankheiten 117 (1) (1944) 76–136. 4

[8] A. P. Association, et al., Diagnostic and statistical manual of mental disorders (DSM-5®), American Psychiatric Pub, 2013. 4

[9] F. Thabtah, Autism spectrum disorder screening: machine learning adaptation and dsm-5 fulfillment, in: Proceedings of the 1st International Conference on Medical and health Informatics 2017, 2017, pp. 1–6. 4

[10] E. Schopler, R. J. Reichler, R. F. DeVellis, K. Daly, Toward objective classification of childhood autism: Childhood autism rating scale (cars)., Journal of autism and developmental disorders. 5

[11] C. Lord, M. Rutter, A. Le Couteur, Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders, Journal of autism and developmental disorders 24 (5) (1994) 659–685. 5

[12] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, M. Rutter, The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism, Journal of autism and developmental disorders 30 (3) (2000) 205–223. 5

[13] J. N. Constantino, C. P. Gruber, Social responsiveness scale: SRS-2, Western psychological services Torrance, CA, 2012. 5

[14] D. P. Wall, J. Kosmicki, T. Deluca, E. Harstad, V. A. Fusaro, Use of machine learning to shorten observation-based screening and diagnosis of autism, Translational psychiatry 2 (4) (2012) e100–e100. 5

[15] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, S. Narayanan, The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody, Journal of Speech, Language, and Hearing Research 57 (4) (2014) 1162–1177. 5

[16] S. Baron-Cohen, S. Cassidy, B. Auyeung, C. Allison, M. Achoukhi, S. Robertson, A. Pohl, M.-C. Lai, Attenuation of typical sex differences in 800 adults with autism vs. 3,900 controls, PloS one 9 (7) (2014) e102251. 5

[17] T. Langdell, Recognition of faces: An approach to the study of autism, Journal of child psychology and psychiatry 19 (3) (1978) 255–268. 5

[18] A. Klin, S. S. Sparrow, A. De Bildt, D. V. Cicchetti, D. J. Cohen, F. R. Volkmar, A normed study of face recognition in autism and related disorders, Journal of autism and developmental disorders 29 (6) (1999) 499–508. 5

[19] J. McPartland, G. Dawson, S. J. Webb, H. Panagiotides, L. J. Carver, Event-related brain potentials reveal anomalies in temporal processing of faces in autism spectrum disorder, Journal of child Psychology and Psychiatry 45 (7) (2004) 1235–1245. 5

[20] E. Pellicano, L. Jeffery, D. Burr, G. Rhodes, Abnormal adaptive face-coding mechanisms in children with autism spectrum disorder, Current Biology 17 (17) (2007) 1508–1512. 5

[21] C. Katarzyna, V. Fred, Limited attentional bias for faces in toddlers with autism spectrum disorders, Archives of general psychiatry 67 (2) (2010) 178–185. 5

[22] S. Weigelt, K. Koldewyn, N. Kanwisher, Face identity recognition in autism spectrum disorders: a review of behavioral studies, Neuroscience & Biobehavioral Reviews 36 (3) (2012) 1060–1084. 5

[23] T. Falck-Ytter, C. von Hofsten, How special is social looking in asd: a review., Progress in brain research (189) (2011) 209–22. 5

[24] K. A. Pelphrey, N. J. Sasson, J. S. Reznick, G. Paul, B. D. Goldman, J. Piven, Visual scanning of faces in autism, Journal of autism and developmental disorders 32 (4) (2002) 249–261. 5

[25] L. L. Speer, A. E. Cook, W. M. McMahon, E. Clark, Face processing in children with autism: Effects of stimulus contents and type, Autism 11 (3) (2007) 265–277. 5

[26] B. Corden, R. Chilvers, D. Skuse, Avoidance of emotionally arousing stimuli predicts social–perceptual impairment in asperger's syndrome, Neuropsychologia 46 (1) (2008) 137–147. 5

[27] T. Falck-Ytter, Face inversion effects in autism: a combined looking time and pupillometric study, Autism Research 1 (5) (2008) 297–306. 5

[28] W. Jones, K. Carr, A. Klin, Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder, Archives of general psychiatry 65 (8) (2008) 946–954. 5

[29] N. Hernandez, A. Metzger, R. Magné, F. Bonnet-Brilhault, S. Roux, C. Barthelemy, J. Martineau, Exploration of core features of a human face by healthy and autistic adults analyzed by visual scanning, Neuropsychologia 47 (4) (2009) 1004–1012. 5

[30] L. Yi, Y. Fan, P. C. Quinn, C. Feng, D. Huang, J. Li, G. Mao, K. Lee, Abnormality in face scanning by children with autism spectrum disorder is limited to the eye region: Evidence from multi-method analyses of eye tracking data, Journal of vision 13 (10) (2013) 5–5. 5

[31] L. Yi, P. C. Quinn, C. Feng, J. Li, H. Ding, K. Lee, Do individuals with autism spectrum disorder process own-and other-race faces differently?, Vision research 107 (2015) 124–132. 5, 11

[32] L. Yi, P. C. Quinn, Y. Fan, D. Huang, C. Feng, L. Joseph, J. Li, K. Lee, Children with autism spectrum disorder scan own-race faces differently from other-race faces, Journal of experimental child psychology 141 (2016) 177–186. 5

[33] S. Wang, J. Xu, M. Jiang, Q. Zhao, R. Hurlemann, R. Adolphs, Autism spectrum disorder, but not amygdala lesions, impairs social attention in visual search, Neuropsychologia 63 (2014) 259–274. 5

[34] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugeson, D. P. Kennedy, R. Adolphs, Q. Zhao, Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking, Neuron 88 (3) (2015) 604–616. 5

[35] A. Klin, W. Jones, R. Schultz, F. Volkmar, D. Cohen, Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism, Archives of general psychiatry 59 (9) (2002) 809–816. 5

[36] J. Van der Geest, C. Kemner, M. Verbaten, H. Van Engeland, Gaze behavior of children with pervasive developmental disorder toward human faces: A fixation time study, Journal of Child Psychology and Psychiatry 43 (5) (2002) 669–678. 5

[37] R. Caldara, S. Miellet, imap: a novel method for statistical fixation mapping of eye movement data, Behavior research methods 43 (3) (2011) 864–878. 5, 7, 30

[38] G. S. Young, N. Merin, S. J. Rogers, 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 12 (5) (2009) 798–814. 5

[39] W. Jones, A. Klin, Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism, Nature 504 (7480) (2013) 427–431. 5

[40] J. Tao, T. Tan, Affective computing: A review, in: International Conference on Affective computing and intelligent interaction, Springer, 2005, pp. 981–995. 6

[41] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, in: IEEE 2011 workshop on automatic speech recognition and understanding, no. EPFL-CONF-192584, IEEE Signal Processing Society, 2011. 6, 39

[42] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402. 6

[43] M.-L. Bourguet, Designing and prototyping multimodal commands., in: Interact, Vol. 3, Citeseer, 2003, pp. 717–720. 6

[44] A. Crippa, C. Salvatore, P. Perego, S. Forti, M. Nobile, M. Molteni, I. Castiglioni, Use of machine learning to identify children with autism and their motor abnormalities, Journal of autism and developmental disorders 45 (7) (2015) 2146–2156. 6

[45] M. Duda, J. Kosmicki, D. Wall, Testing the accuracy of an observation-based classifier for rapid detection of autism risk, Translational psychiatry 4 (8) (2014) e424. 6

[46] J. Kosmicki, V. Sochat, M. Duda, D. Wall, Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning, Translational psychiatry 5 (2) (2015) e514. 6

[47] G. Deshpande, L. Libero, K. R. Sreenivasan, H. Deshpande, R. K. Kana, Identification of neural connectivity signatures of autism using machine learning, Frontiers in human neuroscience 7 (2013) 670. 6

[48] D. Stahl, A. Pickles, M. Elsabbagh, M. H. Johnson, B. Team, et al., Novel machine learning methods for erp analysis: a validation from research on infants at risk for autism, Developmental neuropsychology 37 (3) (2012) 274–298. 6

[49] Y. Zhou, F. Yu, T. Duong, Multiparametric mri characterization and prediction in autism spectrum disorder using graph theory and machine learning, PLoS One 9 (6) (2014) e90405. 6

[50] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, S. Narayanan, Applying machine learning to facilitate autism diagnostics: Pitfalls and promises, Journal of autism and developmental disorders (2014) 1–16. 6

[51] M. Jiang, Q. Zhao, Learning visual attention to identify people with autism spectrum disorder, in: Proceedings of the ieee international conference on computer vision, 2017, pp. 3267–3276. 6, 7

[52] S. Chen, Q. Zhao, Attention-based autism spectrum disorder screening with privileged modality, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1181–1190. 6

[53] J. Bidwell, I. A. Essa, A. Rozga, G. D. Abowd, Measuring child visual attention using markerless head tracking from color and depth sensing cameras, in: Proceedings of the 16th International Conference on Multimodal Interaction, ACM, 2014, pp. 447–454. 6, 39

[54] J. Sivic, A. Zisserman, Efficient visual search of videos cast as text retrieval, IEEE transactions on pattern analysis and machine intelligence 31 (4) (2008) 591–606. 6

[55] M. Rosenblatt, et al., Remarks on some nonparametric estimates of a density function, The Annals of Mathematical Statistics 27 (3) (1956) 832–837. 7

[56] E. Parzen, On estimation of a probability density function and mode, The annals of mathematical statistics 33 (3) (1962) 1065–1076. 7

[57] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Discriminative learned dictionaries for local image analysis, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8. 7

[58] Z. Jiang, Z. Lin, L. S. Davis, Learning a discriminative dictionary for sparse coding via label consistent k-svd, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1697–1704. 7

[59] M. Yang, L. Zhang, X. Feng, D. Zhang, Fisher discrimination dictionary learning for sparse representation, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 543–550. 7

[60] U. L. Altintakan, A. Yazici, Towards effective image classification using class-specific codebooks and distinctive local features, IEEE Transactions on Multimedia 17 (3) (2015) 323–332. 7, 24

[61] S. Zhang, Q. Tian, G. Hua, Q. Huang, S. Li, Descriptive visual words and visual phrases for image applications, in: Proceedings of the 17th ACM international conference on Multimedia, ACM, 2009, pp. 75–84. 7

[62] S. Singh, A. Gupta, A. A. Efros, Unsupervised discovery of mid-level discriminative patches, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 73–86. 7

[63] C. Doersch, S. Singh, A. Gupta, J. Sivic, A. Efros, What makes paris look like paris?, ACM Transactions on Graphics 31 (4). 7

[64] C. Doersch, A. Gupta, A. A. Efros, Mid-level visual element discovery as discriminative mode seeking, in: Advances in neural information processing systems, 2013, pp. 494–502. 7, 16, 24, 25, 26, 27

[65] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, RMPE: Regional multi-person pose estimation, in: ICCV, 2017. 7, 41

[66] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, C. Lu, Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, arXiv preprint arXiv:1812.00324. 7, 41

[67] Y. Xiu, J. Li, H. Wang, Y. Fang, C. Lu, Pose Flow: Efficient online pose tracking, in: BMVC, 2018. 7, 41

[68] D. E. King, Dlib-ml: A machine learning toolkit, Journal of Machine Learning Research 10 (2009) 1755–1758. 7

[69] D. L. Yue Zhao, Yuanjun Xiong, Mmaction, https://github.com/open-mmlab/mmaction (2019). 7

[70] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng, et al., Applying eye tracking to identify autism spectrum disorder in children, Journal of autism and developmental disorders 49 (1) (2019) 209–215. 7

[71] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, P. L. Callet, A dataset of eye movements for the children with autism spectrum disorder, in: Proceedings of the 10th ACM Multimedia Systems Conference, 2019, pp. 255–260. 7

[72] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, F. Saeed, Asd-diagnet: a hybrid learning approach for detection of autism spectrum disorder using fmri data, Frontiers in neuroinformatics 13 (2019) 70. 8

[73] E. Feczko, N. Balba, O. Miranda-Dominguez, M. Cordova, S. Karalunas, L. Irwin, D. Demeter, A. Hill, B. Langhorst, J. G. Painter, et al., Subtyping cognitive profiles in autism spectrum disorder using a functional random forest algorithm, Neuroimage 172 (2018) 674–688. 8

[74] O. Altay, M. Ulas, Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and k-nearest neighbor in children, in: 2018 6th international symposium on digital forensic and security (ISDFS), IEEE, 2018, pp. 1–4. 8

[75] E. Leblanc, P. Washington, M. Varma, K. Dunlap, Y. Penev, A. Kline, D. P. Wall, Feature replacement methods enable reliable home video analysis for machine learning detection of autism, Scientific reports 10 (1) (2020) 1–11. 8

[76] N. A. Chi, P. Washington, A. Kline, A. Husic, C. Hou, C. He, K. Dunlap, D. Wall, Classifying autism from crowdsourced semi-structured speech recordings: A machine learning approach, arXiv preprint arXiv:2201.00927. 8

[77] T. Sorensen, E. Zane, T. Feng, S. Narayanan, R. Grossman, Cross-modal coordination of face-directed gaze and emotional speech production in school-aged children and adolescents with asd, Scientific reports 9 (1) (2019) 1–11. 8

[78] L. Yi, C. Feng, P. C. Quinn, H. Ding, J. Li, Y. Liu, K. Lee, Do individuals with and without autism spectrum disorder scan faces differently? a new multi-method look at an existing controversy, Autism Research 7 (1) (2014) 72–83. 12, 23

[79] W. Liu, L. Yi, Z. Yu, X. Zou, B. Raj, M. Li, Efficient autism spectrum disorder prediction with eye movement: A machine learning framework, in: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, 2015, pp. 649–655. 12, 23

[80] Z. S. Harris, Distributional structure., Word. 14

[81] J. Sivic, A. Zisserman, Efficient visual search of videos cast as text retrieval, Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (4) (2009) 591–606. 14

[82] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, IEEE, 2005, pp. 524–531. 14

[83] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (5) (2002) 603–619. 18, 24

[84] W. Liu, M. Li, L. Yi, Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework, Autism Research 9 (8) (2016) 888–898. 23, 24, 25, 26, 27

[85] N. I. of Health, et al., Autism spectrum disorder fact sheet, Retrieved from. 33

[86] A. S. Nadig, S. Ozonoff, G. S. Young, A. Rozga, M. Sigman, S. J. Rogers, A prospective study of response to name in infants at risk for autism, Archives of pediatrics & adolescent medicine 161 (4) (2007) 378–383. 34

[87] I. Bretherton, The origins of attachment theory: John bowlby and mary ainsworth., Developmental psychology 28 (5) (1992) 759. 35

[88] M. A. Hofer, Psychobiological roots of early attachment, Current Directions in Psychological Science 15 (2) (2006) 84–88. 36

[89] R. L. Grzadzinski, R. Luyster, A. G. Spencer, C. Lord, Attachment in young children with autism spectrum disorders: An examination of separation and reunion behaviors with both mothers and fathers, Autism 18 (2) (2014) 85–96. 37

[90] M. A. Gernsbacher, J. L. Stevenson, S. Khandakar, H. H. Goldsmith, Why does joint attention look atypical in autism?, Child Development Perspectives 2 (1) (2008) 38–45. 38

[91] K. Chawarska, A. Klin, R. Paul, S. Macari, F. Volkmar, A prospective study of toddlers with asd: short-term diagnostic and cognitive outcomes, Journal of Child Psychology and Psychiatry 50 (10) (2009) 1235–1245. 38

[92] D. E. King, Dlib-ml: A machine learning toolkit, Journal of Machine Learning Research 10 (Jul) (2009) 1755–1758. 40

[93] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874. 40

[94] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, arXiv preprint arXiv:1612.02295. 40

[95] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37. 41

[96] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European conference on computer vision, Springer, 2016, pp. 483–499. 41

[97] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1116–1124. 43

[98] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection?, in: European Conference on Computer Vision, Springer, 2016, pp. 443–457. 48

[99] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction, IEEE Transactions on Pattern Analysis and Machine Intelligence. 50

[100] R. Sturner, B. Howard, P. Bergmann, T. Morrel, R. Landa, K. Walton, D. Marks, Accurate autism screening at the 18-month well-child visit requires different strategies than at 24 months, Journal of autism and developmental disorders 47 (10) (2017) 3296–3310. 56