AUTOMATED ASSESSMENT OF NON-VERBAL SOCIAL BEHAVIORS IN EDUCATIONAL CONTEXTS USING DEEP LEARNING FRAMEWORKS

by

Zhang Guo

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Summer 2022

AUTOMATED ASSESSMENT OF NON-VERBAL SOCIAL BEHAVIORS IN EDUCATIONAL CONTEXTS USING DEEP LEARNING FRAMEWORKS

by

Zhang Guo

Approved: _____

Rudolf Eigenmann, Ph.D.
Interim Chair of the Department of Computer and Information Sciences

Approved: _____

Levi Thompson, Ph.D.
Dean of the College of Engineering

Approved: _____

Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Roghayeh Barmaki, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Christopher Rasmussen, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Sunita Chandrasekaran, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Anjana Bhat, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

I am thankful to many people for helping me through out my journey both directly and indirectly. Firstly I would like to express my sincere gratitude to my advisor, Professor Roghayeh (Leila) Barmaki, for not only providing me the exciting research opportunities but also teaching me how to be a researcher during the past few years. She has always been inspiring in guiding me through various academic spheres, encouraging me to constantly surpassing myself, and affirming my achievements. She put enormous efforts into providing all-around support to help me overcome the difficulties in my research and my life. I am grateful for her patience, motivation, immense knowledge, invaluable assistance, scientific advice and many insightful discussions and suggestions. I will treasure this experience as a unique and an enlightening one.

Beside my advisor I would like to thank Professor Christopher Rasmussen, Professor Sunita Chandrasekaran, and Professor Anjana Bhat for serving on my dissertation committee and giving much appreciated and valuable feedback.

I also would like to acknowledge the sponsors of my research, Amazon Research Awards Program, UD Research Foundation, UD College of Engineering, the National Institutes of Mental Health (NIMH, 5R21MH089441-02, 4R33MH089441-03), and Autism Speaks (Grant #8137) for their generous support.

My achievements would not have been possible without the support systems provided by the University of Delaware College of Engineering, and the Computer and Information Sciences department throughout the years. I would like to express my gratitude to Dr. Mary Martin and Graduate College for their tremendous support in my last year of PhD. I would also like to thank the instructors who teach the valuable courses that I took as a student or I assisted as a teaching assistant, and the faculties who provide suggestions, support, and guidance of my research and my PhD program.

I also want to thank all my friends, Jicheng Li, Bilin Sun, Yu Li, Kangsoo Kim, Yan Ming Chiou, Eeshita Biswas, Daniel Gaston, Xiaolu Xu, Fanruo Meng, and our HCI@UD lab members who enlighten me academically and personally during our memorable time together in my doctoral experience. Finally, I would like to thank my family's unconditional help and support.

This thesis would not have been possible without you. Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Non-verbal social behaviors, including gaze movements, facial expressions, and body gestures, help educators and trainers measure students' social interactions and evaluate their learning performance during the educational process. Collecting non-verbal signals between students and teachers/collaborators in a co-located learning setting requires a significant amount of time and effort from data-analysis researchers to manually collect, monitor, and analyze students' behaviors during the learning process. With the rapidly developing educational technologies, it is critical to devise more efficient and reliable tools that can reduce annotation costs and automatically comprehend the students' non-verbal social behavior states in educational environments. Over the past two decades, there has been significant progress in deep learning-based computer vision methods that exhibit a superior capability in visual feature extraction and drive intelligent applications in multiple disciplines without human intervention. The blossom of deep learning provided benefits in big data, computational power, and algorithms.

This dissertation employs computer vision methods with deep learning-based frameworks to automatically extract non-verbal social interaction features from video recordings captured during the learning process. It analyzes students' learning performance based on the detected features in (special) educational contexts. The non-verbal symbols that we mainly focus on include joint attention and mutual attention (overall the social visual behaviors) of two targeted subgroups: university students and children with autism. The deep learning-based frameworks we used in this dissertation are state-of-the-art methods but have not been applied to new domain problems related to education. The research outcomes from our works include the quantifiable,

objective measures and computational models that we developed for social interaction measurements and learning performance analytics as follows.

Firstly, we used an object detection method based on the Mask R-CNN framework to detect/locate the students and their learning tools from the image data collected from a team-based anatomy learning activity. We then proposed a method for quantifying the physical proximity information from students' locations in the activity room and evaluated collaborative actions based on the team's physical proximity dynamics. Despite its strengths for small-group activity analysis, this proximity-based measure cannot handle static/seated position settings, such as standard classrooms. Thus, we investigated other social/collaborative metrics beyond physical proximity, including gaze-related indicators. As the next step, we looked into a gaze point prediction method using the Gaze Following framework for joint visual attention (JVA) measurement and team performance evaluation using the image data from collaborative anatomy learning activity as a test-case study. We found that the JVA frequency of collaborators was a reliable measure in the successful distinction of the study conditions for most educational scenarios. We later introduced an automated social visual behaviors assessment tool by mutual gaze detection method along with an in-house autism dataset collected from a therapeutic intervention in Dr. Bhat's research group. The mutual gaze ratio generated from the detection outcomes was comparable to the social visual behavior score hand-coded by the therapy experts. Lastly, we introduced a social visual behavior analytics approach based on the advanced mutual gaze detection method and the expanded autism dataset. We found that mutual gaze behavior could generate reasonable non-verbal social behavior measures in learning analytics, especially in special education contexts.

Our contribution includes providing multiple automated non-verbal social behavior assessment tools to assist and replace traditional hand-coded annotation; applying computer vision approaches with the state-of-the-art deep learning-based frameworks to solve new domain problems in educational contexts; generating objective non-verbal social behavior indicators for education and social behavior analytics; applying

different learning analysis approaches to evaluate our proposed methods. Our findings have implications for teaching and learning technologies in educational environments, special training, and autism therapy analysis by offering novel assessment tools and analytic approaches for non-verbal behaviors. Beyond the educational contexts, our work can also be applied to scenarios that demand reliable automatic behavior analytics from video/image data involving human-human or human-virtual agent social interactions.

Chapter 1

INTRODUCTION

Non-verbal communication, as a type of interpersonal communication without relying on human language, is an observable, reliable, and essential human-human interaction that people use day to day. Reportedly more than 50% of communication is performed through non-verbal symbols, which can send a strong message regardless of what your words say [7]. For example, smiling when you meet someone shows friendliness, acceptance, and openness; eye contact with high frequency and following the speaker's eye movement shows you are listening carefully; and people are more likely to physically move towards/against those they like/dislike. Non-verbal cues are an essential aspect of human-human relationships, and mastering skills in executing and interpreting them is critical.

More specifically, in academic and educational environments, students use eye movements, facial expressions, body gestures, physical proximity, and more to establish good interactions with teachers, other students, and other academic community members during the learning process, which is a critical skill-set for both learning and teaching process. It is believed that the frequency and quality of non-verbal communication influence academic achievement and student learning behaviors [8, 9, 10]. As an innovative measurement, non-verbal communication can provide scholars with new dimensions to solve educational problems by tracking and measuring the non-verbal symbols between students and instructors during the learning process. Many researchers have studied different non-verbal symbols for communication in normal/special learning settings [11, 12, 13] and the learning benefits they provided [8, 9, 10]. This dissertation aims to investigate the non-verbal social interaction between student-student pairs and student-teacher pairs in both normal and special educational contexts.

Collaborative learning, as a team-based and student-centered educational approach, is an essential educational strategy for teaching and learning. It can also promote students' motivation and enhances knowledge retention via teamwork in both normal [14] and special educational settings [15]. While collaborative learning has been widely introduced and practiced in co-located settings [16, 17, 18, 19, 14] and distributed settings [20, 21, 22, 23], collaboration evaluation remains a challenging task. In the activity-based learning process, learners may get closer during the discussion and cooperation. Thus, physical proximity can be used as an objective measure for collaborative learning analytics in a co-located learning intervention. However, this measurement cannot handle the contexts with fixed positions, such as the classroom settings.

As one of the most critical non-verbal communication skills, the effects of gaze have been studied in terms of its various functions in social interactions [24]. Gaze-oriented cues can be used as a measure of obtaining the cognitive activities of a collaborator, and evidence that students attend to the same object during collaborative co-located learning activities [19, 25]. This gaze alignment behavior is called joint visual attention (JVA [26], see Figure 1.1 as an example). JVA is a powerful predictor of good collaboration among students [27, 28]. Capturing JVA features during the entire collaborative process helps measure the quality of interactions among teams [29].

In special education, individuals and children with Autism Spectrum Disorder (ASD), a developmental disorder with qualitative impairments in social behaviors and communication, have difficulty in identifying, performing, and maintaining such social gaze behaviors [30]. The deflection in social abilities often leads to fear, anxiety, depression, and avoidance of the individuals on the Autism spectrum during the social interactions [31]. The impairment can also lower the quality of interaction,for example, reduction in eye contact [32], interest in social stimuli [33], response to name [34], sharing of interests [35], and social connections with partners [36], further complicating our research. Reportedly 1 in every 54 children is on the autism spectrum in the United States [37], and the number has continued to increase over the past decades. The

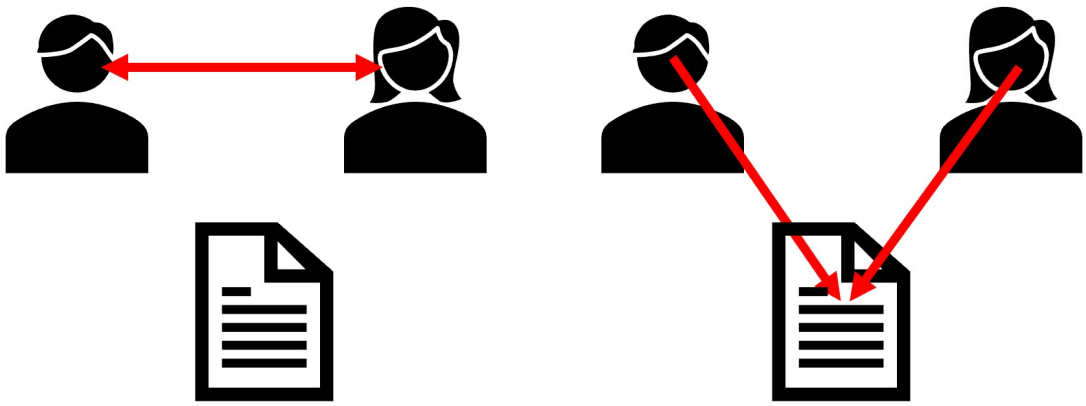


Figure 1.1: Examples for mutual attention (left) and joint attention (right). The red arrows represent gaze directions.

urgency and importance of appropriate therapeutic services create a pressing need for novel therapies and automated assessment solutions that balance cost and effectiveness in autism therapy contexts.

Mutual gaze is considered as a cue for establishing and maintaining successful face-to-face interactions during the special education training process [38, 39, 16] (see Figure 1.1 as an example). In autism therapy interventions, mutual gaze has been used to interpret the social behaviors of children with autism and therapy effectiveness evaluation [40, 41]. However, for social visual behavior analytics in children with autism, manually collecting and evaluating gaze data is quite challenging and requires a huge amount of time and effort from therapists.

Compared with collecting traditional survey data or learning system data for one-time performance evaluation [42], capturing social behaviors during the entire process can reveal critical information about the quality of human-human interactions [29]. Traditional analysis methods such as hand annotations cannot be easily applied to large-scale image- and video-based datasets. Studying the functions and effects of such non-verbal symbols from the camera-captured datasets still requires a significant amount of time and effort from human experts [43, 40, 41]. Therefore, it is important to devise more effective and efficient tools that can reduce human experts' workload by automatically recognizing and analyzing the non-verbal social behaviors of students

during the learning process.

There is a growing interest in implementing wearable devices for non-verbal symbols detection, motivated by advances in hardware technology [44, 45, 39]. However, those devices are expensive to deploy for large numbers of people and are not suitable for naturalistic face-to-face interactions [39]. An alternative and effective solution is to extract and estimate non-verbal features via computer vision methods by standard cameras [1, 46, 47]. Nowadays, computer vision techniques, as the sub-discipline of artificial intelligence (AI) that enables computer systems to interpret the visual world, have been widely applied to people's daily life, including video analysis, object recognition, person detection, and beyond. Compared with the time-consuming hand-coded annotations (manually going through, measuring, and analyzing the social behaviors from images or video recordings), automatically extracting useful features via computer vision approaches provides a more reliable and efficient assessment tool for further data analysis.

1.1 Statement

For purposes of this dissertation, deep learning-based computer vision methods can efficiently extract non-verbal interactive features from images or videos and provide objective measures for learning performance analysis. This dissertation aims to provide automated non-verbal social interactions assessment tools and learning performance analytic approaches that assist or even replace traditional, time-consuming hand-coded annotations provided by subject matter experts. We mainly focus on the non-verbal social behaviors of students in image-based and video-based data, including physical proximity, joint visual attention, and mutual gaze attention. For automatic analysis of visual recordings of social and educational settings, we adopt state-of-the-art deep learning-based frameworks and computer vision methods for domain-specific feature generation and extraction, such as Mask-RCNN [4] for physical proximity analysis of teams during collaborative learning tasks and gaze following [3] and head tracking [5, 6] for team's gaze dynamic understanding. We generated objective measures leveraging

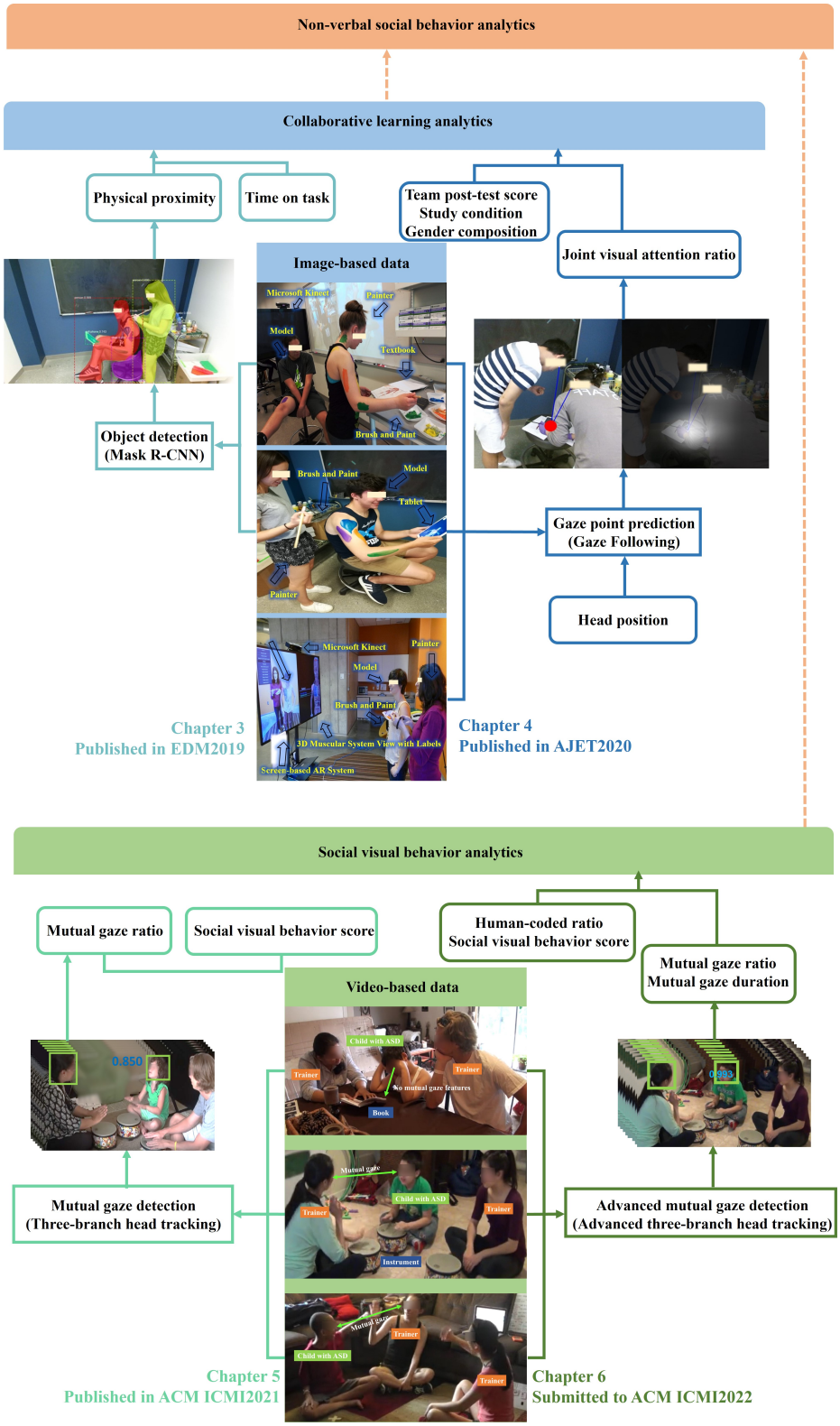


Figure 1.2: Overview of topics covered in this dissertation. Our work shows the feature extraction and analytics towards non-verbal communication from the image-based or video-based data captured from the learning process. Each different colored block shows the research work of a chapter in this dissertation.

framework outcomes to evaluate our methods and analyzed the non-verbal social behaviors using different learning analytic approaches and machine learning models. See Figure 1.2 for topics covered in this dissertation.

1.2 Contributions

Automated non-verbal social behavior assessment tools We provide different automated non-verbal social behavior assessment tools by adopting deep learning-based computer vision methods for non-verbal social behavior extraction. Our methods can assist or even replace traditional time-consuming hand-annotation and assessment by data-analysis researchers.

Innovative applications of computer vision approaches We apply computer vision approaches with state-of-the-art deep learning-based frameworks to solve new educational problems. We adopt Mask R-CNN [4] for object detection and Gaze Following [3] for gaze point prediction to evaluate peer collaboration performance in the co-located collaborative learning intervention. Two different head tracking frameworks [5, 6] are used for mutual gaze detection to assist social visual behavior analytics for children with autism in therapeutic intervention.

Objective non-verbal social behavior indicators We generate objective non-verbal social behavior indicators for learning and social behavior analytics. The non-verbal features extracted by the deep learning-based frameworks, including physical proximity and the frequency of joint attention between collaborators, and the frequency/duration of mutual attention performed by children with autism, are used as objective measures for learning performance evaluation and non-verbal social behavior analytics.

Various learning analytics approaches We apply different learning analytics approaches to evaluate our proposed methods and human learning performance. We used descriptive comparisons, correlation analysis, and regression prediction with various machine learning models to evaluate our proposed methods and investigate the learning performance across group settings, activities, gender, and time period.

1.3 Blueprint of the Dissertation

The remained chapters of this dissertation are structured as follows.

Chapter 2 describes the related work covering the literature reviews on the role of non-verbal communication in education, automated coding of non-verbal data, and the general background for collaboration analytics, social visual behavior analytics, and gaze detection.

In Chapter 3, we provide a new solution to improve the automated collaborative learning analyses using deep neural networks. Instead of using self-reported questionnaires, which are subject to bias and noise, we automatically extract group-working information by object recognition results using Mask R-CNN [4]. This process is based on detecting the people and other objects from pictures and video clips of the collaborative learning process, then evaluating the mobile learning performance using the collaborative indicators. We test our approach to automatically evaluate the group-work collaboration in a controlled study while performing an anatomy body painting intervention. The results indicate that our approach recognizes the differences in collaborations among teams of treatment and control groups in the case study. This work has been completed and published in the Proceedings of the Educational Data Mining in 2019 [8].

In Chapter 4, we introduce an automated team assessment tool based on gaze points and joint visual attention (JVA) information drawn from computer vision solutions. We evaluated team collaborations in an undergraduate anatomy learning activity as a test user study. The results indicate that the JVA ratio is positively associated with student learning outcomes. Moreover, teams who participate in two experimental groups and use interactive 3D anatomy models have a higher JVA ratio and better knowledge retention than those in the control group. Also, no significant difference is observed based on JVA for different gender compositions of teams. This work has been completed and published in the Australasian Journal of Educational Technology in 2020 [39].

In Chapter 5, we introduce an automated mutual gaze detection framework [5],

grounded based on previous works on automated gaze detection, as an effective predictive model for social visual behavior analysis and assessment in autism therapy. To evaluate the proposed gaze prediction framework, we prepare an in-house video dataset that captures social interactions between children with autism and therapists. We estimate the mutual gaze ratio of children using our prediction model, then compare it with the social visual behavior scores from human experts. The results show that our framework provides mutual gaze ratio scores that reliably represent (or even replace) the therapy experts' hand-coded social visual behavior scores by various analysis approaches: descriptive comparisons, correlation analysis, and regression prediction. We report our findings and discuss the implications of the proposed work in the context of visual behavior analysis for children with autism. This work has been completed and published in the Proceedings of the International Conference on Multimodal Interaction in 2021 [48].

In Chapter 6, we introduce a social visual behavior analytics approach by measuring the mutual gaze performance of the children participants in the autism therapy intervention, using an advanced mutual gaze detection framework [6]. Our analysis is based on the expanded in-house video dataset that captures social interactions between children with autism and their therapy trainers. The effectiveness of our framework is evaluated by comparing the mutual gaze ratio derived from the mutual gaze detection framework with the human-coded ratio that therapy experts manually annotated. We analyze the mutual gaze frequency and duration across different therapy settings, training activities, and sessions. We use the mutual gaze-related measures for social visual behavior score prediction with multiple machine learning-based regression models. The results show that our method provided mutual gaze measures that reliably represent (or even replace) the therapy experts' hand-coded social gaze measures and effectively evaluated and predicted children's social visual performance during the therapy training intervention. This work has been submitted to the Proceedings of the ACM's International Conference on Multimodal Interaction in May 2022.

Chapter 7 draws a conclusion to this dissertation and points out research implications, limitations, and potential future directions.

Chapter 2

RELATED WORK

2.1 Non-verbal Communication in Education

A review of the impact of teachers' non-verbal behaviors on the development of students in the classroom is reported by Comadena, Hung, and Simonds [49]. According to these authors, research regarding non-verbal communication has consistently illustrated that the specific non-verbal symbols used by the teachers will have a direct impact on the psychological relationship between teachers and students. They also assert that, in the educational environment, non-verbal communication helps intimacy creation and serves as the foundation of the connection between teachers and students. This connection can have an obvious influence on the overall academic performance of the students. Mackay [50] reports that students are more likely to respond first to the non-verbal body language used by the teachers in the classroom. A teacher's facial expression, gaze, voice, gesture, and movement convey confidence and control, or lack of these. And the changes in the non-verbal patterns can transform the attention of students.

Scholars also examine the impacts of non-verbal communication within different situational needs in the educational environment. Sime [51] and Liu [52] assert that on-verbal communication can be used to reinforce cognitive learning, emotion connection, and classroom management. The research provided by Houser and Frymier [53] demonstrates the importance of non-verbal communication in student development and the need for congruity between verbal and non-verbal symbols provided by the teachers. Non-verbal communication also impacts the development and management of the conflict between verbal and non-verbal cues [54] and within non-verbal symbols [55].

Rupert and Neill [55] note that the teachers needed to effectively recognize conflict and employ non-aggressive responses by controlling their non-verbal behaviors.

Current research on non-verbal communication has extensively focused on the cross-cultural and intercultural context. A report by Helmer and Eddy [56] notes that Non-verbal communication in a culturally diverse educational environment may be misinterpreted by students, which causes conflict or barriers for effective interaction. Le Roux [57] also mentions this issue of cultural competence in non-verbal communication and asserts that teachers need to investigate the patterns of non-verbal communication in a culturally diverse group and implement culturally sensitive communication strategies in order to better understand the specific needs of students and reduce the threat of conflict. In intercultural settings, non-verbal communication is an essential element of successful interactions between partners from different cultures. It is reported by the researchers that, compared with verbal communication, non-verbal massages are less systematized and entirely culturally construed [58].

In recent studies [59, 60, 61], a growing number of non-verbally oriented and methodologically complex researches are more likely to document a wider range of effects using biologically measures and real-time outcomes. Yet, most of this data coding has been conducted manually with professional human coders, typically at around 30-second intervals, and recorded for the presence or absence, frequency, and duration of non-verbal behaviors [62]. However, manual coding of the non-verbal communication data and other audiovisual content is a time-consuming, painstaking, and tedious process. In data collection stages, to ensure accuracy, multiple passes of the same content are always required, and for every hour of content, it takes multiple hours to perform reliable manual coding [63, 64].

2.2 Automated Coding Non-verbal Data

Over the past two decades, numerous computational analytic tools–such as information cascades [65], cluster or classifier analysis for pattern recognition [66], or

opinion mining and sentiment analysis [67]–have been developed and applied to various research questions in communication by the researchers. Given the large volume of research in non-verbal behavior analysis and visual communication, it's also practical to apply advanced computational tools to large-scale data sets of images or videos [47]. Compared to the expensive and time-consuming manual coding of visual data, computational tools can speed up research progress with lower coding costs when they are working [59]. In computer vision and machine learning aspects, these tools have been developed over decades, but their quality and identification rate were not reliable enough for actual applications in the real world until recently when the field made a tremendous increase in accuracy by using artificial deep neural networks [68].

Artificial neural network, introduced in the 19 century [69], is a computational model constructed by a number of internal nodes and their connections, and it resembles the biological neural network in its structure and the way that information is passed between neurons. However, neural networks were not widely used based on the computational complexity and training difficulty at that time. In recent years, the development of GPU-based algorithms and the availability of large-scale training data sets have made neural networks become the most popular, powerful, and reliable machine learning framework [70]. The popularity of neural networks has been enhanced by several famous examples: IBM's Deep Blue beat Garry Kasparov in chess; Google's AlphaGo defeated one of the best players, Lee Sedol, in the game of Go'.

Advanced computer vision and deep learning methods have been increasingly developed and adopted by scholars in the non-verbal communication area to ease the burden of manual coding and widen the scope of analysis [8, 47]. Bantupalli and Xie [71] propose a deep learning framework to recognize American sign language for people with impaired hearing and speech with the American Sign Language Dataset. Talegaonkar et al. [72] create a real-time facial expression recognition model using a deep neural network that can be used for emotion analysis while users watch movie trailers or video lectures. As these examples demonstrate, automated computational methods in computer vision and deep learning can fit into social science frameworks

and enable large scale quantitative analysis of non-verbal data.

2.3 Collaboration Analysis Using Object Detection

Collaborative learning is a widely used education pattern featured by small group interaction and team-based evaluation metrics. Typically two or more participants are assigned in the same group and work for a common purpose which encourages them to learn via teamwork and cooperation [73]. With the computer-supported technology, many researchers even focus on online collaborative learning or long-distance collaborative learning based on the online system, multimedia, virtual reality [74], and mobile social applications [75]. For large number of students in one class, compared with the traditional individual learning or lecture-based learning in the classroom, effective collaborative interaction with peers promotes a positive attitude toward the subject matter for the students and increases student retention.

Using appropriate education technology is one of the main factors of facilitating learning and improving learning performance. To meet the 21st-century students' requirement, the term called "mobile learning" has been introduced by Jacob, and Issac, in 2008, which was based on the usage of mobile tablets for learning purposes [76]. According to the study of Cheung and Haw [77], mobile learning is an advanced strategy to motivate students engage into study and diversify the teaching model.

Learning the terminology and the concept like human anatomy challenges undergraduate students. Time is spent on somewhat inefficient learning by attending lectures, reading textbooks and reviewing 2-D images [78]. Besides using the mobile tablets, many successful efforts were proposed recently to improve active learning. Body painting [79] , clay modeling [80], haptics simulation and models [81], and other models or methods used in education have added playful learning experiences.

The core of analyzing collaborative learning process is to build the cooperation indicators using the data gathered from group working and estimate the quality of the cooperation process [82]. Various approaches for analyzing group working interaction have been proposed. Soller and Lesgold [83] have provided an effective collaborative

learning model as a framework, collected peer-to-peer conversation data using questionnaires and evaluated the active learning skills by computing the value for the following attributes: encourage, reinforce, rephrase, lead, suggest, elaborate, explain/clarify, justify, assert, information, elaboration, clarification, justification, and opinion. In 2007, they provided Hidden Markov Modeling approach to model collaborative learning [84]. Collazos et. al [82] divided the collaborative learning process into three stages and defined 5 of indicators to evaluate the collaborative learning process: applying strategies, intra-group cooperation, success criteria review, monitoring and performance. Later, in 2007 [85], they developed software tools to gather information automatically and used system-based measurement to improve their evaluation framework. With the widespread use of computers, computer supported collaborative work, has reshaped data collection and analytic in collaborative learning settings.

In Chapter 3 we proposed to use object detection techniques for collaborative learning analyses. Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class in digital images and videos [86]. The modern history of object detection goes along with the development of deep learning techniques. Among all deep learning architectures, deep convolutional neural networks [70], widely used as image feature extractor, have the most impact on computer vision tasks. When dealing with the task, several popular deep neural networks stand out due to their high performance. AlexNet [70], based on LeNet [87], opened the new era due to the huge lead in ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The R-CNN approach [88], a natural combination of heuristic region proposal method and ConvNet feature extractor, is able to detect the object and localize the object by bounding boxes. However, the problem with the R-CNN method is incredibly slow computation speed. Two improvements are introduced afterwards, Fast R-CNN [89] and Faster R-CNN [69]. Faster R-CNN uses the same algorithm as R-CNN to extract region proposals. The difference comes from RoIPool (RoI represents Region of Interesting) module which works by extracting a fixed-size window from the feature map and using these features to obtain

the final class label and bounding box. It is effective and end-to-end trainable and the high speed makes it able to do real-time object detection [90]. Mask R-CNN [4] extends Faster R-CNN for instance segmentation by adding a branch for predicting class-specific object mask, in parallel with the existing bounding box regressor and object classifier. The additional mask output comes from the finer spatial layout of the object. Mask R-CNN can be generalized to other tasks as well, such as human poses estimation,without changing the same framework. In this experiment, we use a pre-trained Mask R-CNN with the COCO [91] dataset as our object detection approach.

2.4 Co-located Collaborative Learning Analytics Using Gaze Point Prediction

In the domain of human anatomy learning, different education technologies have recently replaced traditional teaching methods such as lectures, cadavers and text-books. With modern computer-assisted technologies, 3D visualisation methods improve students' performance by allowing them to explore 3D anatomical models on 2D mobile device screens [92]. Mobile-based applications and web-based 3D games have been used as learning tools for the study of human skeletal, muscular and cardiovascular systems to name a few [93, 94]. Virtual reality (VR) and augmented reality (AR) techniques have been adopted into medical education and surgical training fields in recent years [95, 96]. As powerful learning tools, VR and AR engage students in an immersive environment with audio and visual interactions, and stereoscopic 3D models to enhance their learning experience [16, 97, 98, 99]. We evaluated team performance in a controlled study that leveraged modern anatomical content visualisation in 3D with handheld tablet devices and large-scale AR displays.

As the results of the different socialisation processes, gender differences have been discussed by recent studies at the team level. The relationship between gender and collaboration is not uniform, and it varies based on different disciplines and tasks [100, 101]. Researchers have shown that females have better information-processing skills than males during cognitive tests [102, 103]. Females' higher management ability in

collaborative tasks has also been highlighted [104, 105, 106]. Several studies [104, 105, 106, 102, 103] concluded that collaboration performance would be improved with females involved. Other research showed that women often had negative experiences on teams due to gender biases at the technical level, especially in science, technology, engineering and mathematics [107, 108]. Conversely, some reported no significant gender effect. Andersson [109] argued that, although females have better performance on individual memory tasks, no gender effect was found in collaborative tasks. Prinsen et al. [18] noted that females were more likely to collaborate, and males were more assertive in computer-mediated communication [110], and computer-supported collaborative learning (CSCL [111]) settings. In the same 2007 study, Prinsen et al. also acknowledged that different distributions of roles in collaborative learning might change gender contributions. We explored potential gender differences in our anatomy learning study in association with learning outcomes and joint attention measures.

The importance of social interactions during the learning process has been emphasised [112]. Collaborative learning not only helps students to improve teamwork skills but also promotes learning motivation, increases learning experience, enhances brainstorming skills [8, 113] and facilitates their learning performance during team interactions [16, 14]. Consequently, researchers have highlighted that instead of using new learning formats, more attention should be paid to the measurements of and access to collaboration performance [17]. In early attempts to analyse collaborative learning, Soller and Lesgold [83] provided a practical collaborative learning framework and evaluated active learning skills using conversational interaction data collected from surveys. With advancements in CSCL research, machine learning techniques have been used to predict student grades, using support vector machines [114], decision trees [115] and regression [116] to name a few. Those solutions either established effective collaborative learning models or built reasonable standards for evaluating collaboration performance based on single-time solicitation techniques. However, the data used in those models were collected from class attendance, quiz scores, or reports, which may represent only students' one-time or episodic performance during the learning activities.

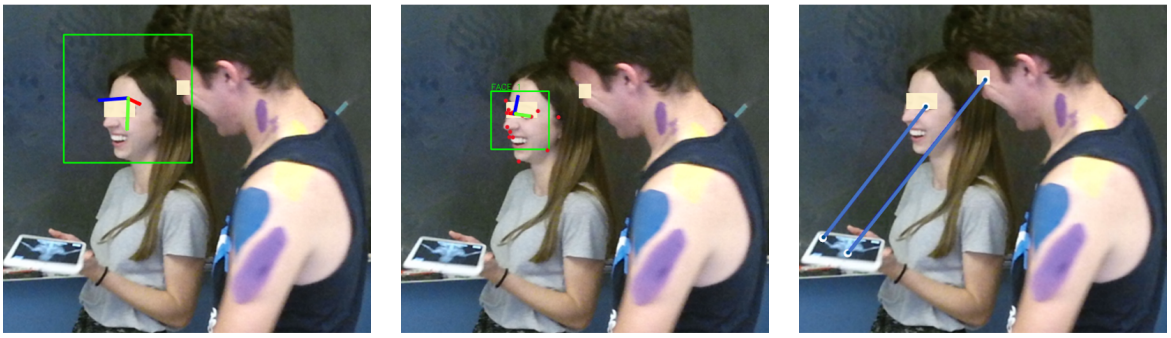
Collaborative learning analytic research has been much of attention mainly in flipped classrooms with collaborative problem-solving activities primarily in science, technology, engineering and mathematics, mediated by computers either in co-located or distributed learning settings. The majority of research for distributed studies lies in CSCL studies; for example, Subburaj et al. [117] presented a collaborative problem-solving model for an educational physics game with 101 teams of undergraduate students. Facial expressions, acoustic-prosodics, eye gaze and task context information were captured in the last minute of the intervention and used as measures for predicting success at solving the game. The combined predictive model of non-verbal cues with language-based features outperformed other predictive models. Also, behavioural cues such as eye gaze [118], head pose [119], prosody and acoustics [120, 121], as well as language [122] have been investigated in collaborative learning analytic for group outcomes including task performance. In co-located learning scenarios, despite their similarity of approaches to distributed settings, physical proximity and movement dynamics in teams were a key factor in the collaboration. In our previous work [8], we used an object detection approach atop image data from a collaborative anatomy learning activity to extract useful proximity features, such as the locations of students and objects in the scene. Research also used multi-modal learning analytic techniques and high-level features from dissimilar sources such as video and sketchpads to discriminate between experts and non-experts in groups of students [123], and understand team performance from physical engagement, satisfaction and individual accountability perspectives [124].

JVA features have been introduced to a broad range of applications, including collaborative search [125], mediated interaction [126], infant-caregiver interaction [127] and training for children with autism [128]. Interest has grown in the use of synchronised eye-trackers to quantitatively measure gaze alignment in various collaborative situations [17, 129, 130, 26]. However, there are challenges in using eye-tracking sensors, including the high cost of the devices, and restricted environmental and calibration settings (e.g., the camera should be precisely in front of the student within a close distance and on top of a specific panel [17]). Image-based computer vision methods – as a

more affordable alternative approach – have also been used for extracting gaze features in previous studies. Using a colour camera, Yücel et al. [131] presented an image-based head pose estimation method for establishing joint attention between an experimenter and a robot. Harari et al. [132] used image segmentation to identify the common gaze target by combining the estimated 3D gaze direction.

There has been an expanding interest in the estimation and reconstruction of human gaze direction from 2D images to identify their activities in the scene using various deep learning frameworks. Gaze following is the task of following people's gaze in a scene and inferring what they are looking at. Compared with eye-tracking and gaze estimation, gaze following not only estimates the gaze direction but also detects the focus point from the image[3]. Patacchiola and Cangelosi [2] proposed a face detector to extract face landmarks and estimate head poses using convolutional neural networks. Marín-Jiménez et al. [1] used head pose detection with implicit pose information to detect human-human interaction in videos. However, those works were limited by the complexity of inputs (massive eye-tracking data [131]; restricted situations (resolution of the image [1]; and field of view (the distance between the camera and students [133]. In the work of Recasens et al. [134], the gaze point of multiple observers in daily scenarios was predicted using deep neural networks and saliency models of attention. Mukherjee and Robertson [135] combined RGB-depth images and multi-modal data to reconstruct 3D head poses and follow gaze direction in images and videos. These studies motivated the work reported here to use a deep learning approach to target gaze alignment features for the novel application of collaborative learning analytic.

In Chapter 4, we were interested in understanding how two students are interacting with one another, or with objects, and following the gazes of multiple observers in a scene. During the preparation stage of our study, we tested various algorithms for gaze feature extraction, including facial landmark detection [2] and head pose detection [1] to predict gaze direction. However, neither facial landmark nor head pose can be detected completely when participants are back to the camera or face downwards, which was not practical for our study (see Figure 2.1a and 2.1b). Hence, we used



(a) (b) (c)

Figure 2.1: Example results from the application of multiple methods on our data set (a) head pose detection [1], (b) facial landmark detection [2] and (c) gaze following method [3].

gaze following method[3] to estimate both the gaze direction and the gaze points to collect human-human interaction information. Further details about our approach are presented in the following section.

2.5 Autism Social Visual Behavior Assessment Using Mutual Gaze Detection

Autism refers to a fast-growing developmental disorder characterized by qualitative impairments in social behaviors and verbal and non-verbal communications [136]. The continued rise in prevalence of autism has promoted extensive research, providing genetic, neurological, and ethnological evidence for autism foundation studies [137, 30, 31]. Children with autism experience deficits in motor control, expressing emotions through facial expressions, social gaze/attention to task, and eventually maybe fearful, anxious, or completely avoid social interactions.

Nowadays, gaze, is one of the most significant, reliable, and observable indicators of non-verbal communication, and has been used in various studies to interpret the social behaviors of children with autism [40, 41]. Converging evidence suggests that impairments in visual attention, including poor social gaze or lack of mutual gaze to caregivers, over-focused attention, or inattention to tasks, are often reported in children with autism from a very young age [138, 139]. Previous studies suggest that mutual

attentiveness is essential for successful dyadic interactions during human face-to-face communication [40, 41]. To this end, tracking the gaze interaction among the children with autism and other people during the therapy/training period can provide therapists with the possibility to investigate children's behaviors and analyze the effectiveness of their treatment.

In therapeutic settings for children with autism, it is important to have an appropriate intervention that best fits the their needs. *Play therapy* is a widely used behavioral therapy for children with autism, in which they are encouraged to express themselves freely through engaging activities. The effectiveness of child-centered play therapy has been investigated and summarized by Hillman [140]. Ware Balch and Ray [141] used a subject design to show the improvement of social and emotional behaviors by child-centered play therapy. Several prior studies [142, 143] in autism-related fields adopted play therapy as an effective evidence-based intervention to promote age-appropriate skills, such as imitation, communication, social attention, and pretend play, through playful activities.

Music therapy has recently become one of the most prominent methods of providing play-based therapy [144, 15, 145]. Among all autism interventions, up to 12% are formed by music-based rhythm interventions [146]. Prior research [147, 148, 15] found that children with autism often had a predisposition for musical stimuli and intact musical perception. Jones [149] addressed the importance of the session frequency and the behaviors of therapy trainers during the music therapy. The results also showed that music-related contexts impacted various forms of development, including verbal and non-verbal communication, social emotion, and motor development in children with autism. Srinivasan et al. [15] conducted a pilot randomized controlled trial and showed that rhythm-based contexts of music therapy could actively promote children's enjoyment and engagement during the therapy sessions, as well as improve their verbal and non-verbal communication and behavioral skills. The effectiveness of the rhythm therapy has been investigated by many other studies as well [150, 144]. Despite the effectiveness of play therapy, which has been quantitatively evaluated using

information-oriented samplings, little research on play therapy video recordings and gaze analysis has been conducted [151, 152, 153]. This research focuses on automatically recognizing gaze interaction from video recordings, thanks to advances in modern computer-assisted technologies.

There is a growing interest in implementing wearable devices for mutual gaze behavior detection, motivated by advances in hardware technology [44, 45, 39]. Ye et al. [44] used wearable gaze-tracking glasses to detect adult-children eye contact. Subsequently, to improve single-frame detection performance, they identified bids for eye contact using a wearable point-of-view camera [45]. However, those devices are expensive to apply for large numbers of people and not completely suitable for naturalistic or face-to-face interactions [39]. An alternative and effective solution could be to estimate gaze direction or extract gaze features using deep learning methods with static cameras [1, 46, 47]. Guo and Barmaki used deep-learning based object detection and gaze following approaches to recognize sharing attention features and analysis collaboration between pairs of students from image-based dataset [8, 39]. In spite of image-based gaze detection tasks, the problem of detecting mutual gaze behaviors in videos was introduced by Marin-Jimenez and his colleagues [1]. They detected human heads in each video frame as a gaze cue and grouped them into tracks. They modeled and predicted pitch and yaw angles of the human heads with a 2D Gaussian Process regression model based on the detected heads, and then computed the mutual gaze score between pairs of heads per frame. Despite the effectiveness, their work only focused on the single frame and lost the consistency of neighboring frames. Palmero et al. [38] extracted mutual gaze features by estimating the eye gaze with a convolutional neural network (CNN, [69]) and used 3D geometry to decide if the gaze ray intersects the head volume of the other person; however, the method was still limited by a controlled scenario that only two people were interacting in the scene. It also required two cameras placed in front of the two target people, and an overlapping visible zone between the two cameras was needed.

Unlike those previous works, the framework by Marin-Jimenez et al. [5], which

we adopted in Chapter 5, automatically derives gaze information from a temporal period of head tracks, whereas previous models have only considered gaze features in a single frame. Compared with other mutual gaze detectors, this deep-learning based framework required only one standard camera located in a fixed position with one or more pairs of heads in the scene. The framework predicted mutual gaze score for all possible pairs in the scene based on their head tracks, even when the head was back to the camera or when the third person cut the gaze ray between the two-side people. Given the importance of mutual gaze analysis in autism therapy addressed above, it is desired to develop a more effective gaze detector, which can consider temporal consistency and relationship in the videos of human activity.

2.6 Autism Social Visual Behavior Analytics Based on Advanced Mutual Gaze Detection

When children are expected to participate in reciprocal behaviors, one of the most immediately observable characteristics of children with autism is their lack of or reduced of eye contact with others or eye-to-face gaze [154]. It is not surprising that visual behavior analytics has received significant attention for its potential to provide a dynamic view of autistic individuals' emotions and their social abilities [155, 156, 157]. Although research into gaze behavior in autism individuals has identified some general patterns, it has also yielded some inconsistent findings: some studies using pictures and videos suggest that autism individuals avoid looking at the eyes [158, 159], whereas others indicate that they have typical gaze patterns as typical controls [160]. It has been illustrated that some of these discrepancies are probably due to the wide spectrum in autism [161], the lack of experimental paradigms for studying eye gaze in real social interaction [162, 163], the difference between various experimental settings and analysis approaches [164, 165, 161].

Recently, there has been a growing interest in visual behavior analytics for face-to-face interaction in order to investigate the gaze patterns of autistic individuals during real social interactions. Volkmar, Fred R., and Linda C. Mayes [164] used a

time-sample technique to collect the frequency of gaze patterns of 20 autistic individuals in educational settings as subjects interacted with staff and engaged in educational activities, and they reported that autistic subjects were more likely to look elsewhere than the matched typical controls and looked less to staff during one-on-one interactions. By accessing the gaze behavior recorded in 20 Chinese children and 23 children with typical development during a face-to-face conversations, Zhao et al. [166] claimed that children with autism looked significantly less at others' mouths and whole faces, and more at the background and they also found that their gaze behaviors varied with the conversational topics. concluded that children with autism looked significantly less at other individuals' mouths and whole faces, while looked more at the background. They also found that their gaze behaviors were varied for different conversational topics depending on how engaging the topic was to the children. Although evidence on interpersonal dynamics is mixed, it is agreed that individuals with autism have difficulties with social dynamics of eye gaze during real-world interactions [32, 148, 159].

Previous work shows that by taking advantage of trace eye-to-face gaze extracted from visual-based data, visual behavior analytics can assess autistic individuals' effective and cognitive behaviors [48, 166, 159]. In order to investigate how individuals with autism visually attend, early studies adopted human raters coded the gaze behavior manually [167]. Typically, manual coding was conducted frame by frame to obtain the gaze allocation, which suffered from the limitations of inaccuracy due to the rater's subjective factors, and extremely time- and labor-consuming [168]. With the advances in hardware technology, wearable devices for gaze tracking can well address these limitations by offering a sensitive and accurate measure of gaze allocation [44, 45, 39, 169, 170]. However, those devices are expensive to widely application and not completely suitable for naturalistic or face-to-face interactions [171, 172, 39].

An alternative and practical solution could be to estimate gaze direction or extract gaze features using deep learning methods with static cameras [1, 46, 47]. We used deep-learning-based object detection and gaze following approaches to recognize sharing attention features and analyze collaboration between pairs of students from

an image-based dataset [8, 39]. In spite of image-based gaze detection tasks, the problem of detecting mutual gaze behaviors in videos was also implemented in recent studies [48, 1, 38]. Marin-Jimenez and his colleagues [1] introduced a mutual gaze detection framework that computed the mutual gaze score between pairs of heads per frame by modeling and predicting pitch and yaw angles of the human heads with a 2D Gaussian Process regression model based on the detected heads. Despite the effectiveness, this work only focused on the single frame and lost the consistency of neighboring frames. Afterward, they provided a new method for mutual gaze detection, which can derive gaze information from a temporal period of head tracks [5]. This work was adopted by us and evaluated by comparing the outcomes with the hand-coded visual behavior measure scored by the therapists [48].

Recently, an advanced mutual gaze detection framework [6] by their teams, which we adopted in Chapter 6, automatically extracted mutual gaze features based on the head tracks with extended temporal dimension head maps and considered multiple consecutive frames instead of single frames in order to reduce the influence of noise, inconsistency, and detection problems. [6]. The head maps also encoded the relative 3D arrangement (depth) of the people in the scene. Compared with other mutual gaze detectors, this deep-learning-based framework required only one standard camera located in a fixed position with one or more pairs of heads in the scene. The framework predicted mutual gaze scores for all possible pairs in the scene. It can also handle the situation that the head was back to the camera or the third person cut the gaze ray between the two-side people.

The Autism Diagnostic Observation Schedule (ADOS) is a gold-standard, semi-structured, standardized assessment for diagnosing ASD, and is an integral part of autism research and clinical protocols for children suspected of having an ASD [173]. The ADOS evaluates social interaction, communication, play, and imaginative play through a series of planned "presses" [174] in the naturalistic social interaction contexts. In each of four developmental- and language-level dependent modules, a protocol of social presses is administered by a trained examiner, and then behavioral items relevant

to ASD are scored with 0 indicating 'typical behaviors' and 1, 2 or 3 indicating 'mild, moderate, to severe ASD symptoms' [175]. A classification indicative of ASD or non-ASD is yielded based on the scores of particular items from the measure. Clinicians or researchers often used this classification as one part of a comprehensive diagnostic process[15, 176, 43].

Chapter 3

COLLABORATION ANALYSIS USING OBJECT DETECTION

3.1 Problem Statement

The past few years have shown that collaborative learning is an effective educational strategy for the educators, who wanted to help those isolate learners join the group interaction, and who realized language skills they taught cannot make meaning without a good language environment and face-to-face practice. Many researches have acknowledged collaborative learning can improve students' learning motivation and increase knowledge retention at both theoretical and empirical level. With the big changes in student populations and the growth in the number of nontraditional learners, small group study format enables students to involve in the discussion, ask questions, and apply their knowledge to solve practical problems instead of individual learning or lecture-based learning.

However, it is undeniable that these benefits only work on the well-performed teams which have efficient collaborative activities during the learning process. Here comes the problem of collaboration learning: how can we find those "at-risk" groups who are likely to fail of group collaboration and need excessive help, and what is the sweet-spot for a well-functioning collaborative team.

Instructors in most universities set group working as part of their course assignments and assign even up to 100% of the final grade for group work. For example, in the affiliated university, several self-reported, and observational measures are used by course instructors to evaluate level of collaboration among teams of students using observation checklists, peer evaluation surveys, focus groups, and the grade of their group products. Though these methods are effective, but there are some constraints

with these approaches, such as no objective measures to automatically evaluate the process of collaboration while students performing a group work. Recently, some scholars either established effective collaborative learning models or built reasonable standards for judging collaborative learning process based on self-reported survey data or collaboration system data. The core of collaborative learning analyses is building indicators based on the participants' performance during the entire learning process.

To encourage students in participation and help them better understand professional concepts and terminologies, different educational technologies and strategies are integrate into collaborative learning. By now, researchers have applied some well-functioned hand-held devices in education and such an education method is called mobile learning. Learners are able to use them to take lecture notes, take online courses, read e-books and have online group discussions. For human anatomy educators, they proposed body painting as an educational strategy, in which those complicate terminology is more visualized and easier to understand than the 2D images on the textbooks. What will happen if we combine these technologies and strategies together and apply them in collaborative learning process simultaneously? To answer this question, collaborative learning analyses are required to be performed on the collected data from both experimental groups and control groups.

What's more, consider using collaborative learning in a large laboratory course, where students are divided into hundreds of small groups. All groups are asked to learn the same objectives, yet they are assigned various learning methods and time spans to complete the learning task. Evaluation of participants' performance and the impact of educational technologies used in collaborative learning require an efficient collaboration data of the entire learning process. Using an automate data collection and analyses method instead of traditional survey will bring significant convenience. Many solutions have been proposed for collaborative learning analyses, such as Regression, Support Vector Machine,and Decision Tree [177, 178]. However, the data for those models were collected from class attendance, quizzes scores, reports, and course views, which could not directly reflect participants' performance during the learning activities. The most

efficient data should be participants' actions and emotions during the process, which could be recorded via images or video clips. Also considering students, typically outnumber instructors in a large scale, instructors could enjoy more convenience collecting data using electronic devices than manually.

Traditional analyses methods cannot be directly applied on the image and video data. Yet object detection methods can help us extract useful features from the raw data. Object detection is a computer technology related to computer vision and image processing that deal with identifying instance of semantic objects of a certain class in digital images and videos [179]. Within object detection domain, there are two main tasks: identification and categorization (also called classification). Image classification is to predict a set of labels to characterize the contents of an input image. Considering classification as a simple supervised problem, the learning module's input is an image and the output is a label of the class of one of the object in the image. The learning module is a binary classifier and it is trained with a couple of labeled examples. Object detection builds on image classification but also allows image localize each object with a bounding box and even a segment. In practice, we need to consider about the size and composition of the training dataset for classification and the variability of each class for identification [180]. Those parameters and detection algorithms vary between different methods.

In this study, our goal of object detection is to detect participants and their learning tools from the image data and then acquire features from participants' locations and finally recognize collaborative actions from the image data. Mask R-CNN [4] (Region-based Convolutional Neural Network) is a Convolutional Neural Network (CNN) model. It has following outstanding features for object detection: (1) it can not only detect the object but also generate a high-quality segmentation mask for each instance; (2) it is simple to train and flexible to use (3) it allows the system to run fast based on parallel heads (4) it can detect multiple objects in one image with high accuracy.

In this work, we evaluate the usefulness of the Mask R-CNN [4] object detection

method in identifying collaborations in a tangible, body painting activity. We first introduce a new team-based anatomy education intervention using electronic hand-held devices, and then compare students' level of collaborative in experimental groups versus control groups using Mask R-CNN [4]. Mask R-CNN [4] trained with the COCO dataset was the method we used for object detection. We will demonstrate the features acquired from the running results of Mask R-CNN [4] could directly reflect students' collaboration performance.

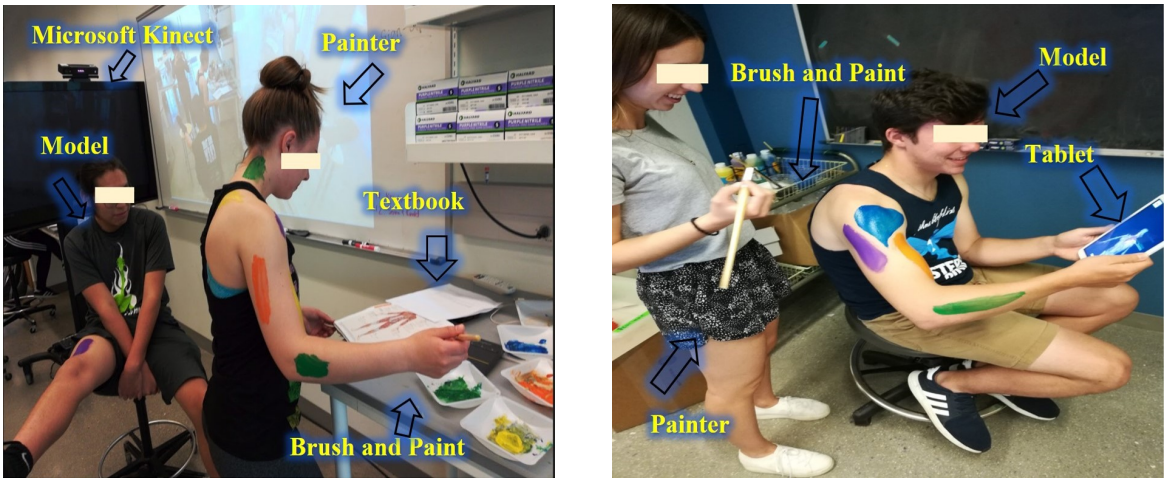
3.2 Materials and Method

In this section, we explain the details of the case study, including the course in which the data were collected, and the specifications of the education technology used in the muscle painting activity and collaborative learning analyses using deep learning method.

3.2.1 Anatomy Learning Intervention

We conducted a between-subjects study while performing an anatomy learning intervention in a large laboratory course of General Biology offered by the Department of Biology at the affiliated University in spring 2018, which was enrolled by over 300 undergraduate students. Students were working in pre-assigned teams for the entire course to complete several anatomy lab exercises, including the muscle painting.

During the muscle painting activity, pairs of students collaborate to paint 12 muscles of their body using painting supplies. The first student plays the role of a model, while her teammate, as a painter, locates the major upper-limb muscles using the human anatomy diagram in the lab manual [181], and then paints her upper limb with painting supplies. Afterwards, students switch roles, and the upper limb painter becomes model for lower limb. The goal of this activity is that students both get the knowledge of anatomy in a collaborative effort. The painting activity was upgraded for experimental group by using mobile tablet devices instead of the textbook for



(a) (b)

Figure 3.1: Case study setup for participants in pairs to complete the painting activity using either (a) textbook, or (b) tablet.

visualizing the musculoskeletal system. Figure 3.1 shows two settings of the study to complete the muscle painting activity.

An online flyer was sent to all students inviting them to participate in the study. The study was approved by the Institutional Review Board (Protocol # HIRB00005021), and oral informed consent was obtained from each participant before the case study commenced. After consent, students completed an online pre-questionnaire (and a pre-test based on their assigned group) individually, and then entered the painting activity room with their preassigned teammates. Each team either used (1) textbook (control conditions), or (2) Tablet (experimental conditions) to complete the painting task. All students completed pre and post questionnaires before and after the intervention.

There was a mobile workstation for each laboratory room to capture videos and photos so that students' performance during the painting activities. The laboratory room was small and clean and participants were asked to stand close to the workstation in order to have appropriate image data.

3.2.2 Dataset

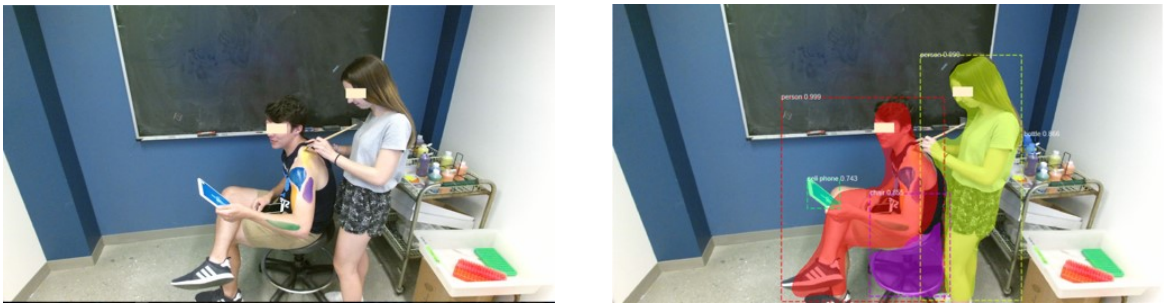
Image data. Image data is the pictures and video clips captured from the muscle painting scene every ten seconds (Figure 3.2a). It showed the entire collaborative learning activities of each group from start to finish. Since we were going to use object detection techniques to detect the people and their behaviors afterwards, only the group members and the tools they used for painting activity were captured in the image data. Approximately 2000 images was captured in total from 33 teams. Based on different time spent on the painting activities, data for each team had about 25 to 200 images. Each image file was also timestamped for further time on task analysis. When choosing the image data for object detection, we focus on how close the team members were during the painting activity and treat such distance as a feature. In this study, we worked on groups of size two.

Survey data. Survey data was collected online, generated by the Qualtrics application, towards all participants before and after completing the painting activities. There were two questionnaires in this study. Pre-questionnaire consisted of demographic questions and a pre-test. Post-questionnaire was composed of questions about participants' user experience during collaboration study, including the preference of being a painter or a model, the level of engagement in the activity, and a post-test. This data set will be used in our future work.

COCO dataset. COCO dataset is an online open source dataset and contains photos of 90 easily recognized common objects categories, including person, chair, desk, bottle, cell phone, book, etc. Over 2.5 million instances are labeled using per-instance segmentation in $328k$ images to aid in precise object localization [91]. COCO dataset can annotate instance-level segmentation mask, and can be used in both image classification and object detection task for both iconic images and non-iconic images [91].

3.2.3 Object Detection Framework

Mask R-CNN [4] approach uses the same Region Proposal Network (RPN) stage as Faster R-CNN to generate bounding box for each candidate object and replaces



(a) (b)

Figure 3.2: Collaborative learning using tablet in painting activity (a) original image data, and (b) object detection using Mask R-CNN [4].

RoIPool by a more accurate module RoIAlign to extract features from each box and show the bounding-box regression and classification [4] (Figure 3.3). By feeding the input image, a CNN feature extractor is able to extract image features which are called feature maps. Then then a CNN RPN will create RoIs which are the candidate object regions generated by RPN and ranked based on their score (how likely is the candidate object region could contain an object). Then the N (Faster R-CNN: $N = 2000$; Mask R-CNN [4]: $N = 300$) regions with the highest scores are kept. Each of them will be warped into fixed dimension by RoIAlign and feed into three parallel branches: two Fully Connected (FC) layers as Faster R-CNN make classification and boundary box prediction and two additional convolution layers to build the mask. The top 100 detection boxes are kept and form a $100 \times L \times 15 \times 15$ tensor, where L represents the number of classes in the training dataset (COCO dataset: $N = 90$), and 15×15 represents the size of each predicted mask. The resized masks and the bounding boxes can be overlaid on the original input image as a transparent layer.

Using a pre-trained neural network by COCO dataset [182], we are able to apply Mask R-CNN [4] to recognize the participants and the painting tools from the collaborative learning activities image data automatically as Figure 3.3. In our implementation, we focused on the following features of the object detection results (Figure 3.2b): image file name, category name, bounding box coordinates and the score for each object. These features are the key points for the collaborative learning analyses and are stored

in an output file in CSV format.

3.2.4 Measures

To evaluate students' performance during the collaborative learning process, we build two indicators: Time on Task, and Level of collaboration in terms of proximity among team members.

Level of collaboration. In this muscle painting activity, collaboration was crucial in terms of how closely the participants work together. Especially close physical distance or proximity, and the amount of time students work together to perform the task was part of the learning process.

We measured this factor using Mask R-CNN [4] approach, and named it level of collaboration. We hypothesize that teams of students who use the tablet, collaborate more with each other, and are within each ether's proximate distance more often in comparison to control group who use textbook. Once again, we want to highlight that our assumption may only fit for tangible work-group activities like body painting with pairs of students, and may not be generalized to other types of group-work activities which need distributed task allocations. Next we describe how we formulated the proximity of students using bounding boxes of study participants.

Overlapping area and ratio. The idea for computing overlapping area and ratio came from how close two participants were during the painting activity. Since there are lots of body contacts between painters and models while collaboration study, their bounding boxes may overlap. Using the coordinates of the bounding boxes of the person, we calculate the overlapping area (if their bounding boxes do not overlap at all, the overlapping area and ratio will be 0) and the whole area for two participants (Figure 3.4). The overlapping ratio is equal to the overlapping area divided by the smaller participant area. The reason of comparing the overlapping area with the smaller participant area is to reduce the bias that one participant may have larger box than the other.

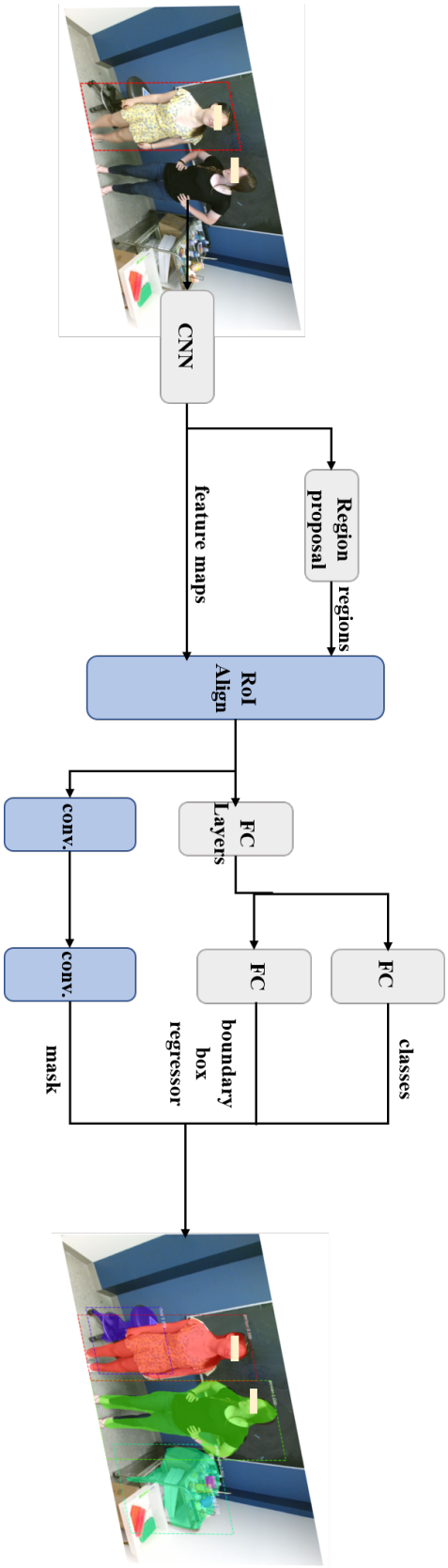


Figure 3.3: The Mask R-CNN framework [4] for object detection.

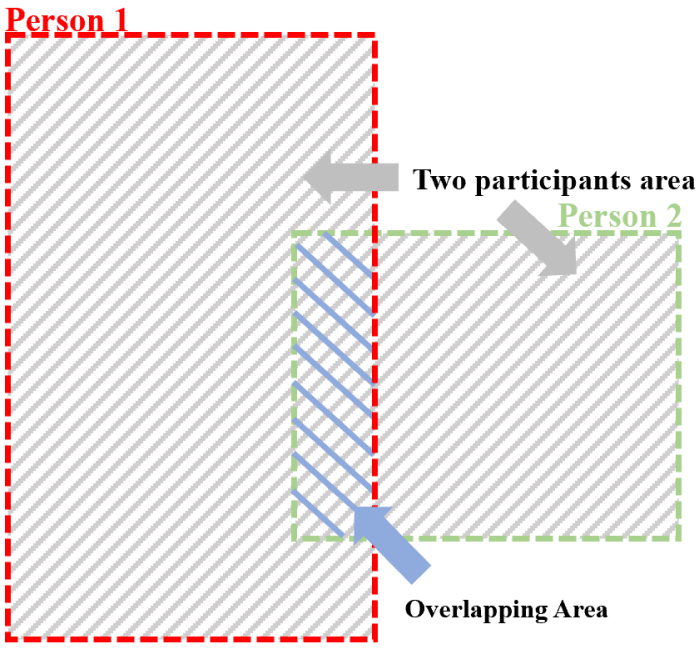


Figure 3.4: The overlapping ratio is equal to the overlapping area divided by the smaller participant area.

Get the percentage of the overlapping rate. For each group, compute the mean of all the overlapping rates in the object detection results.

Time on task. We also calculated the time that students dedicated to performing the collaborative painting task. Previous work [16] shows that increased time on task is a strong predictor of knowledge retention. Un-engaged students just finish the activity in the minimum possible time. In this case, we consider time on task as one of the indicators of student engagement as well. The longer time they spend, the better the collaborative performance they have. This indicator is computed by the file names of the image data and the number of images for each group (time-stamped file information).

Since the treatment group use the mobile learning technology [77] during the muscle painting activity, we hypothesize that teams of students who use the tablet have better collaborative performance than the control group. In comparison to the control group who use the textbook, students in the treatment group collaborate more with each other within proximate distance and they may spend more time on task in

Table 3.1: Summary of the descriptive analytic for two groups of the study.

Group	Level of Collaboration (%) $M \pm SD$	Time on Task (seconds) $M \pm SD$
Treatment	9.23 ± 2.85	631.18 ± 353.75
Control	6.39 ± 1.84	428.75 ± 170.64

Level of collaboration was measured based on the percentage of overlapping area of dyads while working on the task.

the collaborative activity.

3.3 Results

In this section, we report the findings from our case study. A summary of results is reported in Table 3.1 and Figure 3.5.

Participants. Total of 66 (39 Female) students in 33 teams participated in this study. 17 out of 33 pairs were in the treatment group (14 Females), and 16 pairs were in the control group (25 Females).

Level of collaboration. Among 33 teams, we calculated level of collaboration using the aforementioned method, and compared the treatment group versus the control group. The results showed that there was a statistically significant difference between these two groups based on level of collaboration; $F_{(1,33)} = 11.42$, $p < 0.005$, $Cohen's\ d = 1.18$ (large effect size).

Time on task. Similarly, treatment group had a meaningful difference with control group based on time on task $F_{(1,33)} = 4.29$, $p < 0.05$, $Cohen's\ d = 0.72$ (relatively large effect size).

According to the results, collaboration level and time on task had large effect size values and high level of significance which indicated that both two indicators have identified the variations among treatment and control conditions successfully. Based on Figure 3.5, the treatment group had higher means and more able to achieve higher value in both indicators. That means, by using mobile tablets, students were able to get closer to each other and would like to spend more time to complete the task than use textbooks. During the activity, teams with the tablet were more likely to share the

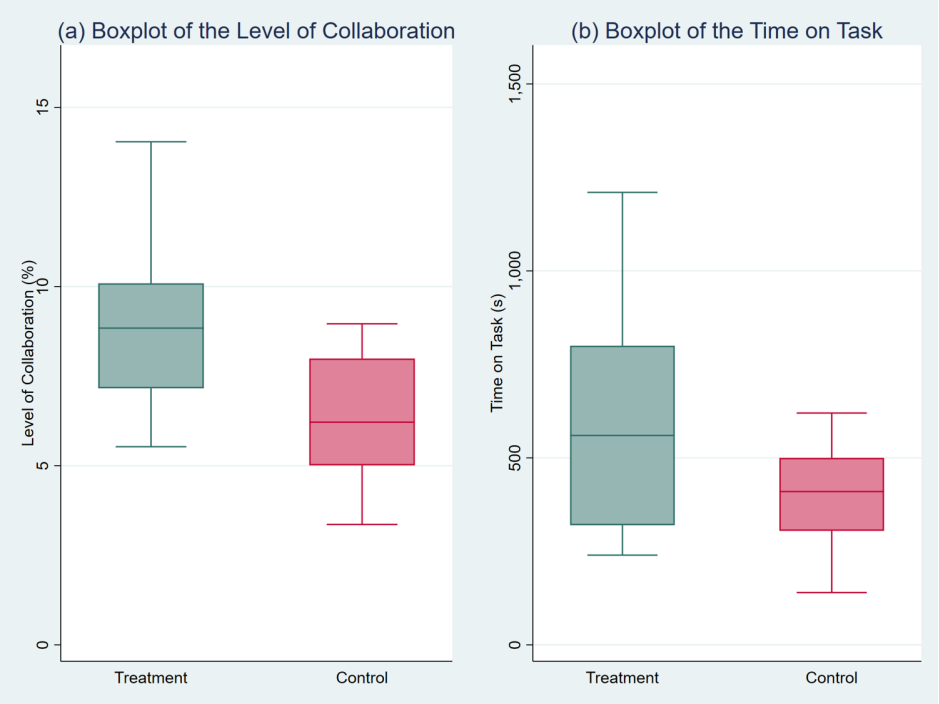


Figure 3.5: The box plots of Level on Collaboration and Time on Task for treatment and control groups in the study.

contents and discuss with teammates in a close distance. Therefore, we could recognize the treatment performed better than the control groups during collaborative learning. Thus, our pre-trained Mask R-CNN [4] approach can provide useful features, and can be used for collaborative learning analysis with acceptable accuracy.

3.4 Discussion and Conclusion

In this chapter, we used Mask R-CNN [4], an object detection approach, to analyze the quality of collaboration. We then evaluated the approach with two collaborative indicators related to team's proximity and time on task. The results showed that our approach was capable of recognizing differences in the level of collaboration among students in treatment versus control groups. Both the time on task and level of collaboration could successfully distinguish the differences between two groups of the study.

The focus of this project is understanding collaboration on muscle painting

activity. Some aspects of our work could be improved in the future. We plan to use our survey data as another dimension to perform collaboration evaluation. We also aim to hand-annotate part of our collected data based on the object categories we need, including painting brush, and re-train our model on the entire data set. We will work on other collaborative features, such as facial expression detection, emotion detection, head, and body pose estimation, and joint attention estimation to better understand the level of collaborations among teams.

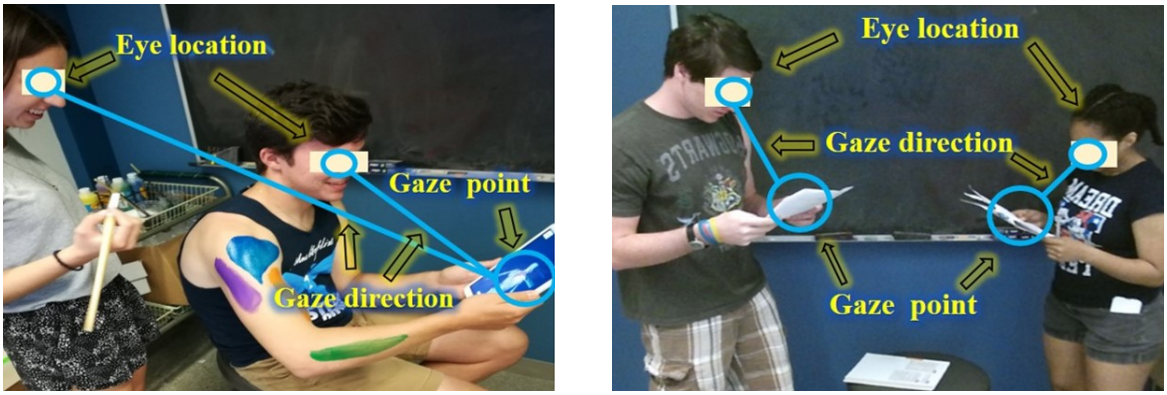
Chapter 4

CO-LOCATED COLLABORATIVE LEARNING ANALYTICS USING GAZE POINT PREDICTION

4.1 Problem Statement

Collaborative learning is an essential educational instrument for teaching and learning. As a team-based and student-centred educational practice, it promotes student motivation and enhances knowledge retention via teamwork and cooperation [14]. While collaborative learning has been introduced and practiced in co-located settings [16, 17, 18, 19, 14], as well as distributed settings[20, 21, 22, 23], measuring and evaluating collaboration still remains a challenge. Fairness of group work distribution [183], rationality of collaborative conditions [184] and automatism of process analytics [115] are some of the core issues that need to be considered during collaborative learning analytics, especially in relatively large teams [185]. Understanding gender effects in collaboration dynamics and investigating best learning practices in teams are also crucial aspects of collaborative learning analytics, which is the focus of this chapter.

Gaze-oriented cues can be used as a means of obtaining information about the cognitive activities of a collaborator, and there is evidence that students look at and point to the same object during collaborative co-located learning activities [19, 25]. This gaze alignment is called joint visual attention (JVA [26])—see Figure 4.1a, for example. JVA is a strong predictor of successful collaboration among students [27, 28]. Compared with collecting traditional self-reported survey data or collaboration system data for one-time performance evaluation [42], capturing gaze alignment during the entire collaborative process with the JVA features can reveal more valuable information



(a) (b)

Figure 4.1: Examples of the gaze-following method in our study: (a) with JVA feature: students' gaze points converge on the tablet; (b) without JVA feature: students look at their own notes.

about the quality of interactions among teams [29]. New technologies provide innovative methods to extract and measure JVA features. A growing number of researchers have taken advantage of sensor-based eye-trackers to objectively measure gaze features during various social interactions, especially for co-located collaborative tasks [17, 19]. However, despite the provision of highly accurate data from eye-tracking devices, these sensors are usually highly priced and may introduce limitations for educational study settings, specifically those needed to be carried out in classrooms and not in research labs. For example, the calibration might be a time-consuming process or the study activity is needed to be completed within the limited tracking range of the sensors.

With the emergence of different deep learning techniques, the gaze tracking problem can be approached differently. Using deep neural networks to track the gaze features from a sequence of 2D images or videos is practical and robust in understanding and interpreting student behaviours in human-human and human-object interaction [3, 134]. For example, when two students are looking for a path from the library to the gym on a single university map, by following their gaze direction, we can easily find out if they are sharing the same information and we can predict whether they will pick the same path. Compared with eye tracking and gaze estimation, the gaze following method [3, 186, 187] not only estimates the gaze direction but also predicts the gaze

point from the image without the need for specialised hardware (e.g., head-mounted camera, infrared light source) and obtrusive gaze calibration procedure. Figure 2.1c shows the application of gaze following method on the image captured from our test user study.

In this chapter, we introduce a computer vision-based solution for team performance evaluation using mutual gaze point predictions, along with a collaborative anatomy learning activity as a test user study for our approach. We recorded collaborative activity sessions as a sequence of images with a colour camera. For collaboration analysis, we first tracked team members' gaze directions and the focus objects during the activity using the gaze following method[3] with a deep neural network framework. We then extracted the JVA features of the teams and analysed them with other collected data, including post-test scores and demographics information related to team gender composition. This study hypothesised that students who share mutual gazes during the activity—for example, those teams with higher JVA values—obtain higher scores in their post-activity knowledge tests as well, since they engage more in collaborative tasks. We were also interested in understanding if JVA values are varied significantly in different study conditions and different gender compositions of teams for collaborative learning.

This chapter is organised as follows: We introduce details of the intervention, dataset, the gaze following method and our proposed assessment measures for collaboration. Then, we present the findings from our test user study and discuss further implications.

4.2 Materials and Method

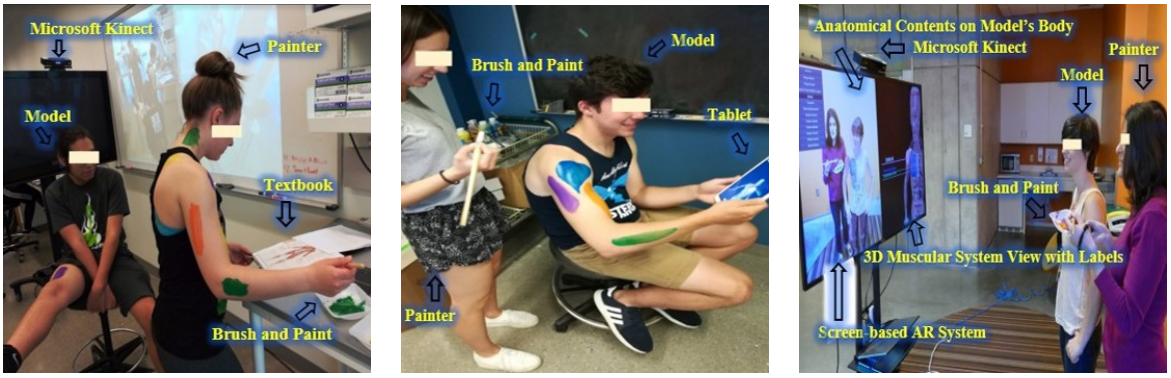
4.2.1 Collaborative Muscle Learning Intervention

We conducted a between-subjects study of collaborative muscle learning intervention in a laboratory course (General Biology) as part of an undergraduate premedical program at Johns Hopkins University. A total of 301 students in 138 teams participated in the original study; we selected the data from a subset of teams with

two members ($N = 60$, 30 teams) as our test data set. Several measures were considered for selecting this subset of the data set. Based on collaborative task distribution, recorded data from teams with more than two members, those in different activity rooms, those with students under 18 years of age and those with incomplete questionnaire data were excluded from our study. Students worked in teams to complete a muscle painting activity [16, 188] as part of their required laboratory activities. They were expected to identify and paint the major muscles of their body using one of the learning instruments (textbook, tablet or AR) and washable painting supplies. The first student played the role of a model, while their teammate, as a painter, located the major upper-limb muscles with the aid of their laboratory manual [188] or other digital devices and painted the model's upper limb. Afterward, students switched roles, and the upper-limb painter became a model for the lower limb. The goal of the learning activity was to ensure all students could gain knowledge of anatomy in a collaborative effort. See Figure 4.2 to learn more about the intervention details.

As briefly mentioned, our study had three different settings based on instrumental tools. Students in the control group used textbooks as their learning tools. In experimental group I, instead of a textbook, students used our in-house interactive app on the tablet as a 3D musculoskeletal visualising system. Experimental group II used a screen-based AR system—also developed internally—where students could see themselves with augmented anatomy visualisations on a large display [16]. The knowledge base information presented in all instrumental tools was identical to mitigate potential confounding factors related to student workload and learning. There was also a mobile workstation inside the laboratory room to capture snapshots from students during learning activities. Figure 4.2 shows the three study conditions of the learning activity.

The study was approved by the Institutional Review Board (# HIRB00005021) in May 2018, and oral informed consent was obtained from each participant student before the study commenced. After consent, students entered the activity room with their teammates and completed the task. All students completed both pre- and post-activity questionnaires and knowledge tests.



(a) (b) (c)

Figure 4.2: Study conditions for students in pairs to complete anatomy painting intervention using (a) a textbook, (b) an interactive app on the tablet, or (c) a screen-based AR system.

4.2.2 Dataset

Surveys. Using the Qualtrics application, survey data was collected from all students individually after they completing the anatomy painting activity. The survey consisted of demographics information, usability questions and a post-test about the human muscle system.

Image training data. We adopted GazeFollow [134], the large-scale gaze-following data set for training from Recasens et al.'s study. This benchmark data set included $130,339$ people and $122,143$ images in total with gaze points inside the image.

Image test data. The test data set consisted of $4,646$ images collected from 30 teams (pairs) of students during the collaborative learning activity in the three conditions (10 teams from each condition of the textbook, tablet and AR, totalling images from 30 teams). Images were captured every 10 seconds, and each image file was timestamped. The resolution of each test image was 2560 pixels. Images with camera difficulties or additional individuals in the scene were discarded.

4.2.3 Gaze Following Framework

To extract shared gaze features from the images, we needed to estimate the students' gaze direction and focus point in the scene. Thus, we applied a two-stage

gaze following approach[3] on our test data set. This method was very suitable for our project since it was capable of detecting the gaze direction from the head image and predict the potential gaze point along the gaze direction, via deep neural networks. The gaze following approach and its underlying network architecture is shown in Figure 4.3.

The gaze following framework was inspired by the human behaviour of gaze following[3]. First, a gaze direction was estimated from the gaze direction pathway. In the gaze direction pathway, the resized head image (224×224)—image sizes are listed in pixels hereafter—was fed into the convolutional neural network ResNet-50 [189] for feature extraction. Then, the head features were concatenated with head position features encoded by one fully connected layer for gaze direction estimation. A coarse gaze direction was predicted as the vector output and then encoded as multi-scale gaze direction fields. The gaze point was assumed to be in the gaze direction or line of sight. Next, the multi-scale gaze direction fields were combined with the scene contents (224×224) and fed into the heat map pathway for heat map regression using a feature pyramid network [190]. The heat map (56×56) represented the probability distribution of the gaze point, and the point with the maximum value of the heat map represented the probable gaze point of the scene.

Lian et al. [3] claimed that their gaze following approach outperformed other existing methods in gaze point prediction. Compared with state of the art[134], Lian et al.'s method decreased 23.68% of the Euclidean distance error for gaze point on the GazeFollow data set. We chose Lian et al.'s gaze following model because it managed to simulate the gaze following behaviour of a third person view. Furthermore, Lian et al. trained this model robustly on a large data set by using the heat maps for focus point prediction. Figure 2.1 highlights its strengths over other existing solutions.

The gaze following output shown in Figure 4.4a visually draws a blue gaze line on the original image for each individual in the scene. The blue line is initiated at the eye location and terminated at the predicted final gaze point (line of sight). The highlighted regions in the corresponding heat map—Figure 4.4d– 4.4f—represent the predicted gaze points where the students are looking. The output also marks the coordinates of

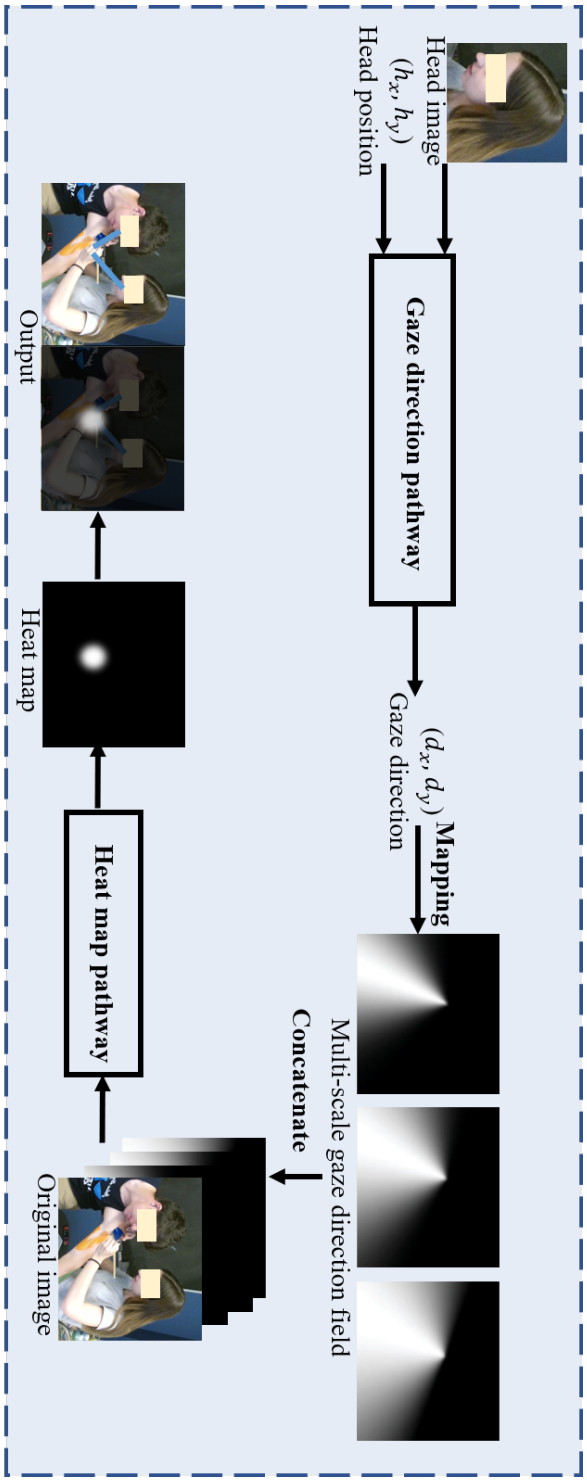
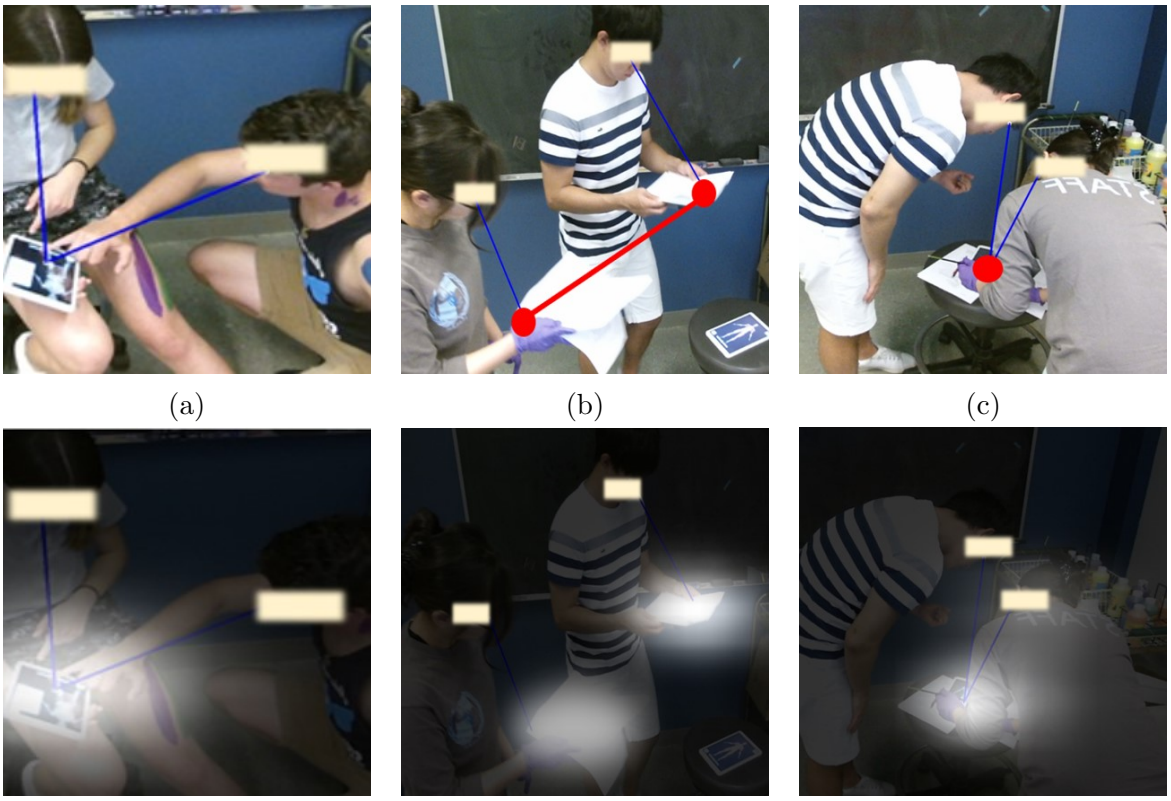


Figure 4.3: The network architecture for the gaze following method [3] atop our collaborative study image frames. Using the heat map, we can predict the gaze point convergence (focus point) of students in the collaborative activity.

each gaze point—which are used in our approach as a collaboration metric. We were interested in the automatic recognition of joint or mutual gaze visual attention among students in every image sequence during the collaborative task. Further details about the JVA feature analysis as a collaboration measure are presented in the following section.



(a) (b) (c)

(d) (e) (f)

Figure 4.4: Gaze following results for three sample frames: (a) gaze directions with blue lines; (b) output without the JVA feature (Euclidean distance between the gaze points of students is greater than 100 pixels; (c) output with the JVA feature (Euclidean distance between the gaze points is smaller than 100 pixels); and (d–f) heat maps associated with the gaze points.
While two distinct sheets of papers are predicted as gaze points for team members in (e), (d) and (f) are examples of the convergence of visual attention on the tablet device; thus, the Is_JVA variable is true in (d) and (f).

4.2.4 Measures

We analysed team performance and collaboration based on objective measures related to joint attention, knowledge retention, study conditions and gender composition of the teams. These evaluation measures of collaboration are described as follows.

JVA ratio. JVA represents the shared focus of two or more individuals and plays a key role in collaboration prediction [191]. In this work, we defined the JVA ratio for each team based on the frequency with which the two students shared gazes during the collaborative activity, divided by total image frames captured from the team—a normalised measure of JVA based on the total activity frames of the teams. Since there was a lot of cooperation between painters and models during the learning activity process, they needed to maintain joint attention most of the time: while painting, discussing and looking at the learning materials. For example, when the painter was painting, both painter and model may have looked at the same location, the active painting region. When students needed to find the muscle's correct location, the painter and the model may have shared the screen of the interactive app on the tablet or the AR view to zoom in on the 3D musculoskeletal system.

We used Euclidean distance between the gaze points detected by the gaze following method for automatic identification of the JVA in each image frame (Is_JVA was a Boolean variable per each frame; it was set to false by default). Based on image size and resolution, we recognised the JVA, or mutual gaze feature, and set Is_JVA as true, if that distance was smaller than 100 pixels. For each team, the JVA ratio was computed by the total number of frames in which the Is_JVA variable was true, divided by the total number of frames. Some examples with and without JVA recognition are shown in Figure 4.1 and Figure 4.4. We were interested to learn if a higher JVA ratio is also associated with better learning outcomes.

Team post-test score. The key objective of this collaborative learning activity was to enhance the anatomy knowledge retention of students. All participants needed to independently—not with the assistance of their peers—locate and label five muscle names in a diagram of the human musculature in the post-test; thus, individual test

scores ranged between 0 and 5 with discrete values. Since we used the average of post-test scores per team and named it team post-test score, the team post-test score was still in the same range, but non-discrete values were also observed in the data set.

Study conditions. As mentioned earlier, there were three different conditions or settings for our muscle painting study. Students in the control group used textbooks as their learning tools. The experimental groups either used a tablet or the screen-based AR system to complete the task. We wanted to investigate the differences in team performance based on these three conditions and the two groups.

Gender composition. Students were preassigned randomly to teams to complete the muscle painting activity. There were three possible gender compositions per pair of students: male pair, female pair and mixed pair. We were interested in evaluating the gender effects in collaborative learning and investigating if any significant variability of JVA ratios and knowledge tests was present in female-female, male-male and male-female (mixed) pairs of students. As females had a higher enrolment rate in the General Biology lab—which was a common pattern in premedical programs across the nation [192, 193]—our study participants were also mostly females with a total of 38 out of 60 study participants (out of the 30 teams, half of them were female-only teams, eight mixed and rest, seven, were male-only teams). This does not mean gender imbalance in the data since the sample population is a representative subset of the target population in premedical programs [192, 193].

4.3 Results

In the following, we report descriptive and inferential results from our test user study. In particular, we looked at the JVA ratio—an automatically generated measure based on our proposed framework using deep neural networks—in association with our evaluation measures. Table 4.1 summarises the descriptive statistics for pairs of students in each study condition, including the number of teams, mean values and standard deviations for JVA ratios and team post-test scores.

Table 4.1: Summary of JVA ratio and team post-test score with different instrumental tools

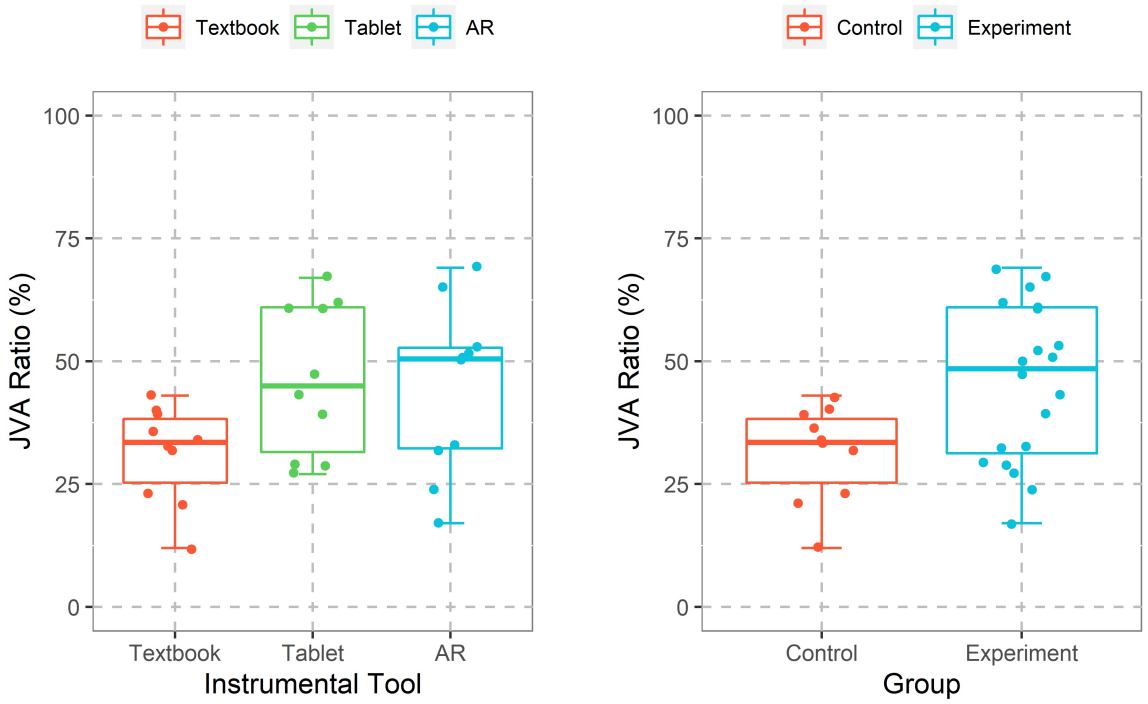
Group	Condition (Instrumental tool)	Observation (teams)	JVA ratio (%)	Team post-test score
		n	$M \pm SD$	$M \pm SD$
Control	Textbook	10	31.30 ± 9.73	1.15 ± 0.95
	Tablet	10	46.50 ± 15.43	2.35 ± 1.03
	AR	10	44.60 ± 17.28	2.35 ± 1.42
Experiment	Combined (Tablet & AR)	20	45.55 ± 15.97	2.35 ± 1.20
Total	Textbook & Tablet & AR	30	40.80 ± 15.59	1.95 ± 1.25

JVA: joint visual attention; M:mean value; n: number of teams; SD: standard deviation; Team post-test score is in [0–5] range; JVA ratio (%) is in [0–100] range.

Participants. We analysed data from 60 participants (38 females) in 30 teams. All of these students were enrolled in the undergraduate premedical program at Johns Hopkins University. There were 10 teams for each condition—textbook, tablet and AR. Knowing that tablet and AR conditions were part of the experimental group, we had 20 teams in the experimental group and 10 in the control group. Data from teams with a larger size, those in different activity rooms, those with students under 18 years of age and those with incomplete data were excluded in this study.

JVA ratio. JVA ratio was the percentage of the time teams had joint attention during the learning activity. Although no significant difference between the three study conditions and the JVA ratio was observed, the p value was very close to the critical value of α ($F_{(2,27)} = 3.26$, $p = 0.054$, ns—ns stands for statistically non-significant). Interestingly, the JVA ratio of the two experimental groups of tablet and AR ($n = 20$, $M = 45.6$, $SD = 15.97$) was significantly higher than those in the control condition who used textbook ($n = 10$, $M = 31.3$, $SD = 9.73$) and this finding was statistically significant with a large effect size ($F_{(1,28)} = 6.65$, $p < 0.05$, $Cohen's\ d = 1.00$). Table 4.1 and Figure 4.5 provide additional information about the JVA ratio distribution across all study conditions and groups.

Team post-test score. A significant difference based on team post-test scores was observed among study conditions of textbook ($M = 1.15$, $SD = 0.95$), tablet ($M = 2.35$, $SD = 1.03$) and AR ($M = 2.35$, $SD = 1.20$), ($F_{(2,27)} = 3.64$, $p < 0.05$, $r^2 = 0.16$,

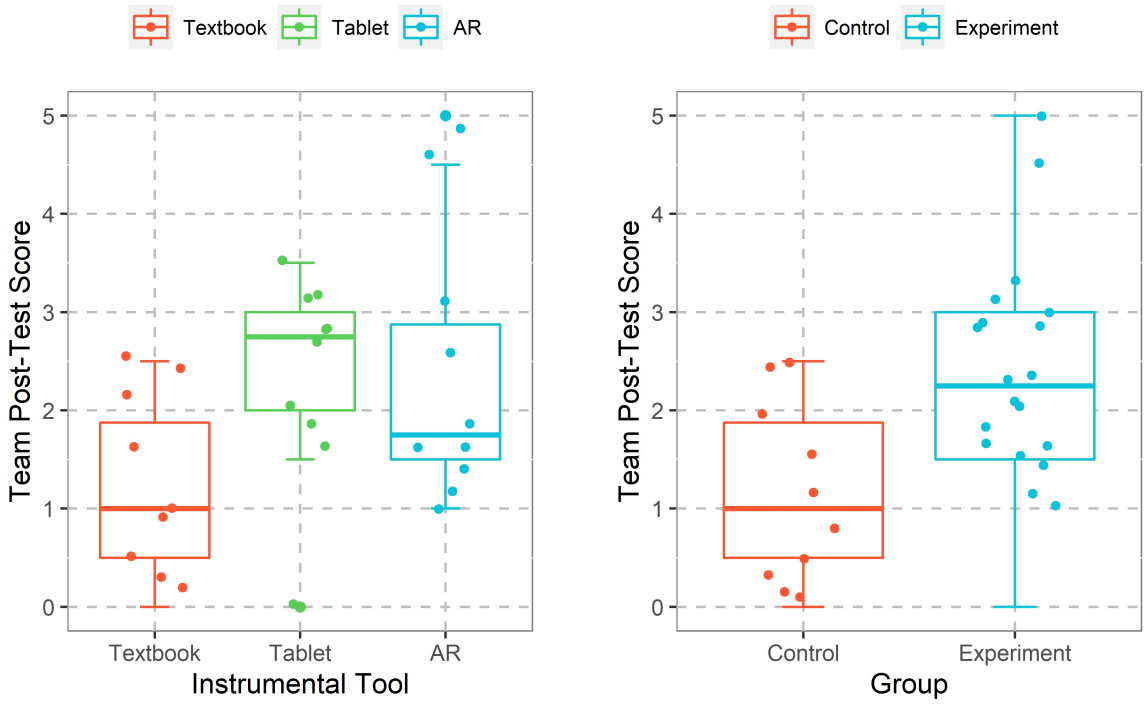


(a) (b)

Figure 4.5: The boxplot with observed data points for JVA ratio across (a) different study conditions (instrumental tools) of textbook, tablet and AR, (b) two groups of control (textbook) and experiment (tablet and AR). JVA ratio was significantly different between control and experimental groups.

medium effect size). Post-hoc comparisons indicate that pairs of textbook and tablet, and textbook and AR conditions were different from each other based on differences in the means. Similarly, the team's average post-test score from the two experimental groups of tablet and AR ($M = 2.35$, $SD = 1.20$) was significantly higher than those in the control group ($M = 1.15$, $SD = 0.95$), and this finding was statistically significant with a large effect size ($F_{(1,28)} = 7.56$, $p < 0.05$, $Cohen's\ d = 1.06$). See Figure 4.6 to learn more.

JVA ratio and team post-test score. We also measured the association between JVA and team post-test scores using the Pearson correlation coefficient. The Pearson correlation measure indicates a significant positive linear association [194] with a strong relationship between the JVA ratio and team post-test scores ($r^2 = 0.50$,



(a) (b)

Figure 4.6: The boxplot with observed data points for team post-test score across (a) different study conditions (instrumental tools) of textbook, tablet and AR, (b) different groups of control and experiment.

$F_{(1,28)} = 9.33$, $p < 0.005$, $r^2 = 0.25$) (large effect size)). The scatter plot drawn from data is shown in Figure 4.7. This finding shows that JVA features are strongly associated with learning outcomes, such as post scores. Points on the scatter plot closely resemble a straight line with a positive slope, which shows that post-test scores increase with higher JVA ratios. Therefore, the team with a high frequency of sharing gazes is more likely to achieve better outcomes in the post-test.

Gender composition. We recorded the gender composition for each of the 30 teams from the survey data and investigated the gender effects on collaborative learning during the activity (see Table 4.2 and Figure 4.8 to learn more). Overall, mixed pairs (eight teams) achieved the highest JVA ratio ($M = 47.0$, $SD = 15.46$) and the best learning outcomes from post-test scores ($M = 2.50$, $SD = 1.28$), but this variability was not statistically significant ($F_{(2,27)} = 1.29$, $p = 0.29$, ns). Moreover, no

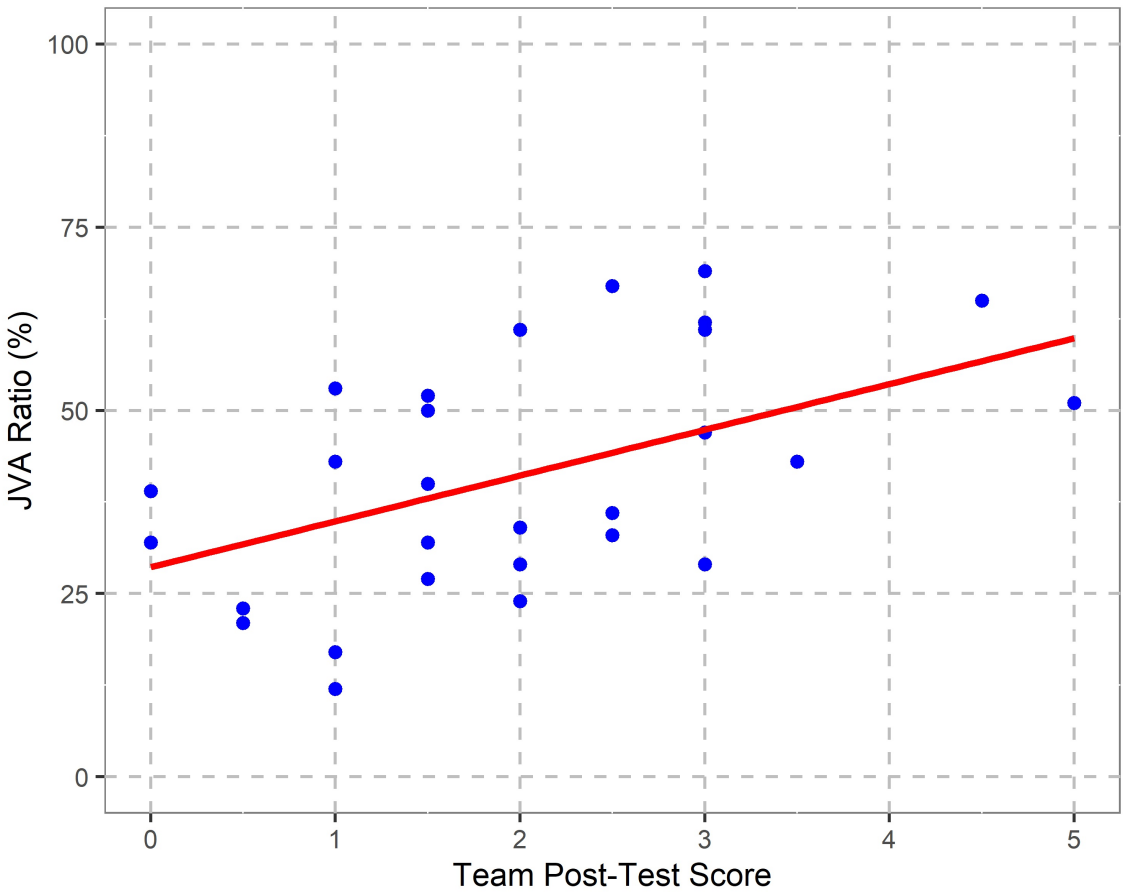


Figure 4.7: The scatter plot of JVA ratio with team post-test scores. The Pearson correlation and its underlying regression model indicate a significant positive correlation between JVA ratio and team post-test score.

significant difference was observed based on gender composition in teams for JVA ratios ($F_{(2,27)} = 1.10$, $p = 0.35$, ns).

4.4 Discussion

JVA ratio. Capturing gaze alignment during the collaborative process with the JVA criterion can reveal valuable information about the quality of interaction among teams [28, 29, 191, 129, 127]; however, not many studies have investigated computer vision-based approaches to better measure and capture it in co-located team-based learning interactions. In this study, we introduced a novel assessment tool for automatic team performance evaluation using mutual gaze information using the gaze

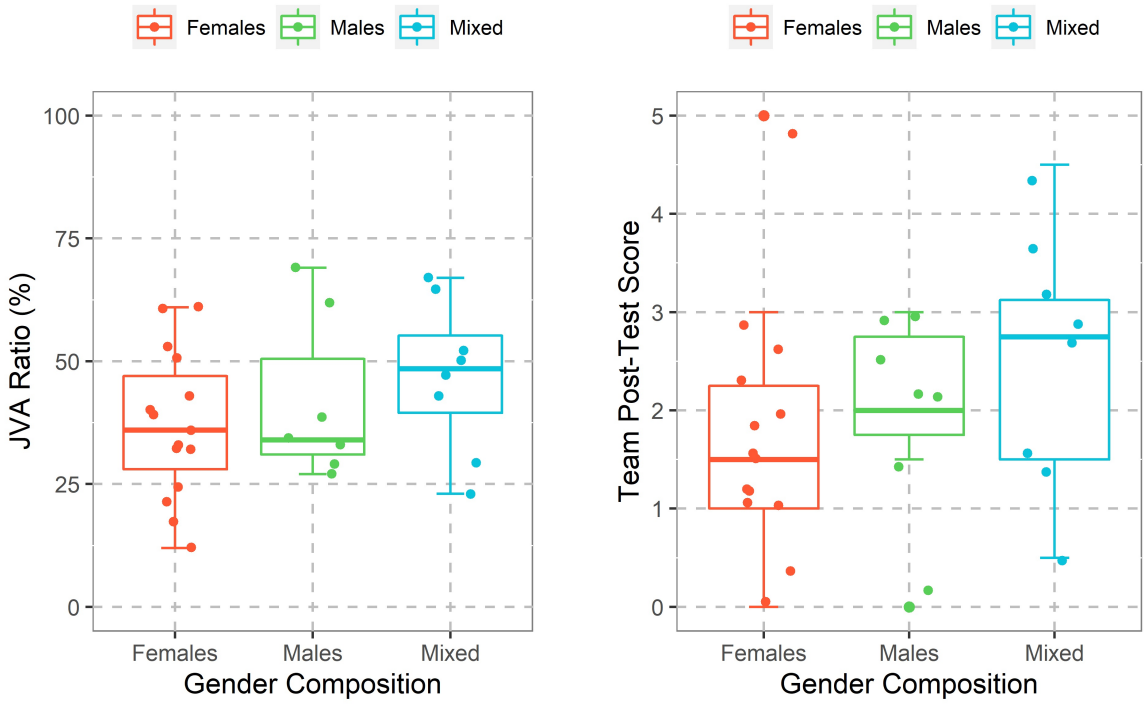
Table 4.2: Summary of JVA ratio and team post-test score with different gender conditions

Gender composition	Observation(teams) n	JVA ratio (%) $M \pm SD$	Team post-test score $M \pm SD$
Females	15	37.00 ± 15.05	1.63 ± 1.29
Males	7	41.86 ± 16.72	2.00 ± 1.04
Mixed	8	47.00 ± 15.46	2.50 ± 1.28
Total	30	40.80 ± 15.59	1.95 ± 1.25

JVA: joint visual attention; M: mean; n: number of teams; SD: standard deviation; Team post-test score is in [0–5] range; JVA ratio (%) is in [0–100] range.

following method[3]. Compared with other methods using traditional one-time performance evaluation [42] or high-cost eye-tracking devices [129], our method was able to automatically extract JVA features during the whole learning process with a simple colour camera. We also investigated the effectiveness of our JVA method in a test user study. Results show that the JVA ratios of the two experimental groups of tablet and AR were significantly higher than those in the control group, who used the textbook. Our findings are supported by research based on gaze information from student users that looked at e-textbooks as a potential alternative learning tool [195], although that research was limited to individual learners and not teams.

Team post-test score. Post-test scores indicate student achievement from the learning activity [196]. In this study, we set up three different study conditions by using different instrumental tools for an anatomy learning activity. Right after the activity, post-test scores were collected from students using a survey completed individually, and the team post-test score was calculated as the average of team members' individual test scores. Team post-test scores of the two experimental groups of tablet and AR were significantly higher than those in the control condition, who used the textbook. Furthermore, our research on collaborative learning analytics conducted with 288 students in May 2017 [16] also showed that higher test scores were achieved from experimental groups who used the AR system. These findings are in agreement with previous studies in anatomy education, which highlighted the potential of using evolving technologies such as mixed and AR for enhancing student learning and outcomes



(a) (b)

Figure 4.8: The boxplot with observed data points across teams with different gender compositions: (a) JVA ratios, (b) team post-test scores. No significant difference was observed in the study for JVA ratios nor post-test scores for different pairs of students. Among these 30 pairs or teams of participants, there were 15 female-female, seven male-male and eight mixed pairs.

in anatomical science education [95, 197, 96].

JVA ratio and team post-test score. There was a significant positive linear association with a strong relationship between the JVA ratio and the team post-test score. This is in agreement with the hypothesis of our study: students who shared mutual gaze with their teammates for a longer time on the learning task were more likely to obtain higher scores in their post-activity knowledge tests. This finding agrees with previous research on the positive effects of sharing a gaze on learning outcomes, including imitation and socio-cognitive performance [198, 199]. Features like hand distance, speed and face count have been used as high-level features for collaborative learning analytics [124], and they seem to have potential in practice-based learning. Our findings

from the JVA and team performance outcomes also show the potential of proximity-based, behavioural measurements in co-located or practice-based collaborative learning analytics.

Gender composition. There were three possible gender compositions per pair of students: females, males and mixed pairs. About half of the participants were in female-only teams, which is also a common gender enrolment rate in life sciences and premedical programs [192, 193]. Even though mixed teams had slightly higher JVA ratios and better learning outcomes, no significant difference was observed based on the gender composition of the teams neither for the JVA ratio or for the team post-test score. These findings are consistent with previous studies that noted no gender effect in life sciences studies [18, 109].

In addition, we found that compared with control groups using text and 2D anatomy models from the textbook, the students in both experimental groups had higher JVA ratios and better knowledge retention by interacting with 3D models on the tablet screen or AR system. Specifically, in teams with the screen-based AR, students could easily collaborate and locate specific muscles with high accuracy projected on top of their own bodies. This outcome was also highlighted in a recent meta-analysis study as collaborative learning being the most beneficial approach in any AR interventions [200]. Our study also provides further evidence that 3D visualisation technologies increase students' engagement and improve their knowledge retention in human anatomy learning [97, 99, 197, 92].

4.5 Conclusion

In this chapter, we introduced an automated team assessment tool based on gaze points and JVA information extracted by computer vision solutions. The results from a pilot user study indicate that experimental teams who interacted with 3D digital learning tools had a high frequency of JVA and better knowledge retention outcomes than those in the control group. We also investigated the association of user study gender composition effects on JVA ratios and team test scores. We found no

significant difference for JVA ratios nor post-test scores for different teams of students with varied gender composition.

This work was a preliminary study to automatically assess team collaboration with computer vision techniques. Like any other project, there is room for improvement. The focus was to understand collaboration on co-located situations, and shared gaze features were identified during the post-analysis process. There was a subset of collected data; since the main objective of this chapter was to first understand the dyadic interactions of students, we excluded larger teams. In future work, it will be ideal to evaluate the effectiveness of adopting our method to other collaborative learning scenarios. In further planned research, a data set with team sizes larger than two would better illustrate our idea and validate our findings. Furthermore, we plan to work on multiple computer vision techniques by combining multiple image-based features, such as facial expression recognition, emotion recognition and head and body pose estimation, along with joint attention estimation to more comprehensively interpret collaboration dynamics. The findings from this work have implications in educational technology and collaborative computing by offering a novel assessment tool for team collaborations based on gaze information.

Chapter 5

AUTISM SOCIAL VISUAL BEHAVIOR ASSESSMENT USING MUTUAL GAZE DETECTION

5.1 Problem Statement

As one of the most critical non-verbal communication skills, the effects of gaze have been studied in terms of its various functions in social interactions [24]. *Mutual gaze* (Figure 5.1), as one of the essential aspects among the functions of gaze, is considered as a cue for establishing and maintaining successful dyadic interactions in social communications [38, 39, 16]. Individuals with Autism Spectrum Disorder (ASD), a developmental disorder with qualitative impairments in social behaviors and communication, have difficulty in identifying, performing, and maintaining such gaze behaviors [30]. The lack of this ability often leads to individuals' fear, anxiety, depression, and avoidance during the social interaction [31], and can lower the quality of interaction, which also makes autism-related information assessment in our society more complex, e.g., reduced eye contact [32], reduced interest in social stimuli [33], lack of response to name [34], and insufficient sharing of interests [35]), and lack of social connections with partners [36].

Reportedly 1 in every 54 children is on the autism spectrum in the United States [37], and the number has continued to increase over the past decades. The urgency and importance of appropriate therapeutic services creates an urgent need for novel therapies that must also be tested for their efficacy. *Play therapy* is a form of behavioral therapy that employs "play" as the basis of the intervention and aims to improve children's social and emotional skills, problem-solving, and verbal and non-verbal behavior within natural settings, and include music/rhythm therapy [15, 176]. However, studying the effects of such therapies still requires a significant amount of

(a) (b) (c)

Figure 5.1: Sample scenes captured from two therapy groups in our ASD dataset: (a) Standard therapy group with reading activity, and Play therapy group with (b) drumming, and (c) singing activity. Mutual gaze features are shown in (b) and (c), while no mutual gaze for child is detected in (a). This chapter aims to automatically capture mutual gaze attentiveness of children with autism, and develop a predictive model for their social visual behavior.

time and effort on the part of data-analysis researchers to collect, monitor, and analyze the behaviors of the children with autism [43]. Given the fast-growing number of children on the autism spectrum, it is timely important to devise more effective and efficient tools that can reduce data-analysis researchers' burden by reliably and automatically recognizing and analyzing the children's gaze attention state during the therapy sessions.

In this chapter, we introduce a deep learning framework, which adopts a state-of-the-art mutual gaze detection method [5], to predict and evaluate mutual gaze behaviors of children in autism therapy interventions. The framework takes the video records as input and predicts the score of the detected mutual gaze as the outcome to automatically detect gaze interaction between therapy trainers and children in the videos. Here, the mutual gaze score is an important measure to identify the gaze behavior and represent the ability of the children to exhibit social behaviors. We demonstrate the effectiveness of the developed framework by comparing the outcome from the framework with the social visual behavior measure hand-coded by therapy experts as ground truth.

The remainder of this chapter is structured as follows. Section 5.2 as the materials and method section describes play therapy interventions with children participants

and our in-house ASD dataset, our deep learning methodology for mutual gaze detection, and the measures for evaluation of our methodology. Section 5.3 presents our results and findings, that is discussed in Section 5.4. Section 5.5 concludes the chapter and informs future research directions and research implications.

5.2 Materials and Method

In this section, we describe the details of our proposed framework for automatic mutual gaze detection and the dataset, which we collected and used for the framework evaluation. First, we introduce different therapeutic interventions with children with autism from our video dataset and describe the details of this in-house ASD dataset. We then describe surveys to collect participant profiles in the dataset and their social, communication, and functional abilities. Finally, we describe our mutual gaze detection framework using a three-branch deep learning approach and introduce our measures for the framework evaluation.

5.2.1 Autism Therapy Interventions and In-House Data Collection

In pursuit of developing an automatic mutual gaze detection framework, particularly for the autism therapy efficacy, we collected a number of video clips of different autism therapy interventions through a multi-session human-subjects study [15].

In a randomized controlled trial design, a total of 16 children with autism and their trainers participated in the intervention. The children were recruited through flyers posted online and onsite in local schools, services, and self/parent advocacy groups, and were randomly assigned to one of the therapy settings. The study was approved by the university Institutional Review Board # 637082-12.

Starting with the first week, children completed pre-test sessions to determine their social, communication, and functional abilities. Then, they had two sessions every week for an eight-week period. Each session included different activities with trainers and the child participant, and it was videotaped with a standard camera. The camera was located at a fixed position towards the child. For each therapy session,

two trainers interacted with the child to complete an embodied creative activity in the child's home environment setting. One trainer introduced the details of the activity and provided guidance for different types of activities. The other trainer was a buddy to the child and practiced all the activities with the child during the entire therapy session. All the trainers were pediatric physical therapists and graduate students/faculties and were trained by autism therapy experts. Finally, in the last week, children completed post-test sessions to determine changes in their social, communication, and functional abilities (totaling ten weeks of intervention, including two weeks of pre- and post-test sessions and eight weeks of the autism therapy activities).

In this chapter, we particularly focused on the participants' visual behaviors during interactions with the trainers. For this reason, we selected a specific subset of video data in the two therapy groups:

- **Standard therapy**: In the Standard therapy group, children engaged in a table-top reading activity (see Figure 5.1a), which promoted academic and social communication skills [15, 201]. During the therapy session, the child was guided to follow the instructions, take turns with trainers to read an age-appropriate book, answer several book-related questions, as well as spontaneous expressions. Session themes included reading books about people, their things, food, transportation, etc.

- **Play therapy**: Participants were engaged in either a drumming activity involving beat keeping/imitation games or a singing activity involving hello or action songs (see Figures 5.1b and 5.1c). During the therapy session, the child was guided to follow the instructions, copy the trainers' movements, and beat the drum to the musical beats or sing a hello/action song along with hand clapping and waving gestures. The themes of this session included various rhythms and musical components, such as start and stop, steady beat, turn-taking, slow and fast, and soft and loud.

In the play therapy sessions, one trainer was seated opposite to the child and the buddy sat next to the child, while in the standard therapy sessions, both trainers were seated next to the child around the table. While both groups engaged in social communication activities such as eye contact, turn-taking, and non-verbal/verbal communication; the

Standard therapy group mainly focused on tabletop reading, whereas the Play therapy group also focused on gross motor skills [63, 64].

In-house ASD dataset. We prepared an in-house ASD dataset for our mutual gaze detection framework based on a subset of the collected video data from the study described above. The data were collected in practical in-the-wild settings where therapists conducted sessions without knowing our framework, which made the balanced samples difficult. For our dataset, we only included videos with annotations by therapy experts—this excluded a volume of videos, which do not have annotations. Six children's video records with annotations were further excluded due to low resolution (2 children records), camera occlusion (2 children records), and incomplete data records (2 children records). Each record in our data included the video clips from sessions 1, 8, and 16, which covered the start, middle, and last sessions of the therapy. Non-annotated videos from other sessions (13 video clips for each child record) were excluded from the analysis. After post-processing, our ASD dataset had a total of 7.5 hours of video from children in 30 video clips (12 for the Standard Therapy group, 18 for the Play Therapy group) with ten children participants (Standard Therapy: 4, Play Therapy: 6). Each video clip was 25 fps with approximately 15-minute length. We used the ASD dataset as the input of our mutual gaze detection framework to calculate the participant's mutual gaze score during the therapy session for further analyses.

Participants profile surveys. Before each child's participation, a parental consent, a Social Communication Questionnaire (SCQ [202]), and a clinical psychology evaluation were completed to confirm their eligibility through different surveys and clinical assessments. The survey data included not only the demographic information of children with autism, including age and sex, but also different physical and social skills, e.g., level of functioning and verbal abilities. The age range of the ten children (3 females) included in our ASD dataset was between 5 and 12 years. The level of functioning of each child was based on their level of independence during daily living skills and was scored on a range from 1 to 3 (1: low functioning or highly dependent, 2: medium functioning or moderately dependent, and 3: high functioning or highly

Table 5.1: Participant profiles for our ASD dataset.

ID	Therapy Group	Function	Verbal	Age	Gender
Child 1	Standard	1	1	12	F
Child 2	Play	1	1	5	M
Child 3	Standard	2	3	6	M
Child 4	Play	1	3	10	M
Child 5	Standard	2	3	8	M
Child 6	Play	2	2	12	M
Child 7	Play	2	3	7	F
Child 8	Play	2	3	7	F
Child 9	Play	2	2	12	M
Child 10	Standard	3	4	5	M

Function: level of functioning skills is scored in range from 1 (low) to 3 (high).
Verbal: level of verbal skills is scored in range from 1 (non-verbal) to 4 (high-verbal).

independent), and the level of verbal skills represented how appropriate/well the child spoke words and sentences, which was scored on a range from 1 to 4 (1: non-verbal or no words used, 2: low-verbal or speaking few words, 3: verbal or using phrase speech, and 4: high verbal or speaking complete sentences). Table 5.1 shows the therapy expert's ratings of the children's functional and verbal levels as well as their sex and ages.

5.2.2 Mutual Gaze Detection Framework

Our primary goal in this research is to develop an a practical computational framework that can automatically detect mutual gaze behaviors in the interactions between the therapy trainers and the children participants. This would be critical for the therapists to understand the state of children with autism and prepare appropriate therapy plans. Here, we describe our development for the mutual gaze detection framework, which adopted a state-of-the-art three-branch head tracking framework [5] on our dataset to extract mutual gaze features between all possible pairs of people from the videos in our ASD dataset. The architecture of the framework is shown in Figure 5.2.

To extract mutual gaze features from videos, we need to track all people and estimate their gaze directions from the scene. Since the head pose (i.e., position and

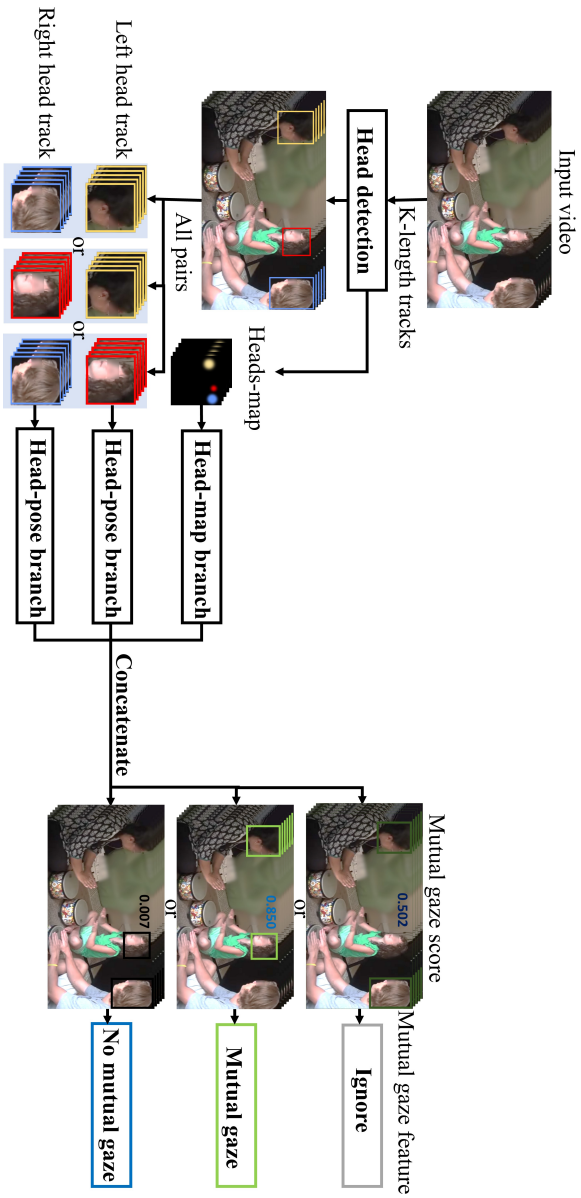


Figure 5.2: Our deep learning architecture for mutual gaze detection, adopted from the three-branch head tracking framework [5] on therapy videos from the ASD dataset. Using mutual gaze scores from the framework outcome, we can recognize mutual gaze features between the child-trainer pairs (trainer-trainer pairs are ignored). The light green bounding boxes are shown with a high mutual gaze score; the dark green/black bounding boxes are shown with a low mutual gaze score, which represents no mutual gaze feature is detected from the scene.

orientation) is considered as a good estimation for the eye gaze while there is no guarantee that the eyes are fully visible in the collected scenes [1], our framework focuses on detecting the head poses of all the people from the frames. A deep CNN is adopted to classify whether each pair of people was looking at each other in the frame tracks. The framework has three input branches, including two head-pose branches and one head-map branch. Two head-pose branches have a tensor of K RGB frame crops of size 64×64 pixels encoded the head sequence of each person of the target pair. The head-map branch is a 64×64 map with 2D Gaussian detected relative heads positions in the central frame of the K-frame track. Head tracks for the head-pose branches are generated by the online linking algorithm [203] using the head detection results from Single Shot Multi-box Detector (SSD [204]). The head-map branch not only encodes position information for the two target persons but also for other persons in the scene, in order to detect the case where the third person cuts the gaze ray between the two-side people. Each head-pose branch is followed by four 3D convolutional layers, and the head-map branch is followed by four 2D convolutional layers. After applying L2-normalization, the outcome embedding vectors from different branches are concatenated and further processed by a fusion block, which includes a fully connected layer with an alternating dropout layer and a Softmax layer to determines a confidence score on whether the pair of people are looking at each other or not, i.e., the score for the mutual gaze. This framework applies to all pairs of simultaneous head tracks in the video clips.

The outcome of the framework consists of the central frames with detected mutual gaze scores for all possible pairs. In each central frame of the track, there are bounding boxes for the heads of each pair with the light or dark green color based on the confidence score (see Figure 5.3, for example). The score shows how likely the pair of target people are looking at each other. In other words, the higher the score is (the lighter the color of the bounding boxes is), the higher likelihood the people look at each other. According to the study conducted in the original paper [5], this framework achieves the state-of-the-art results on the TVHID dataset [205], which

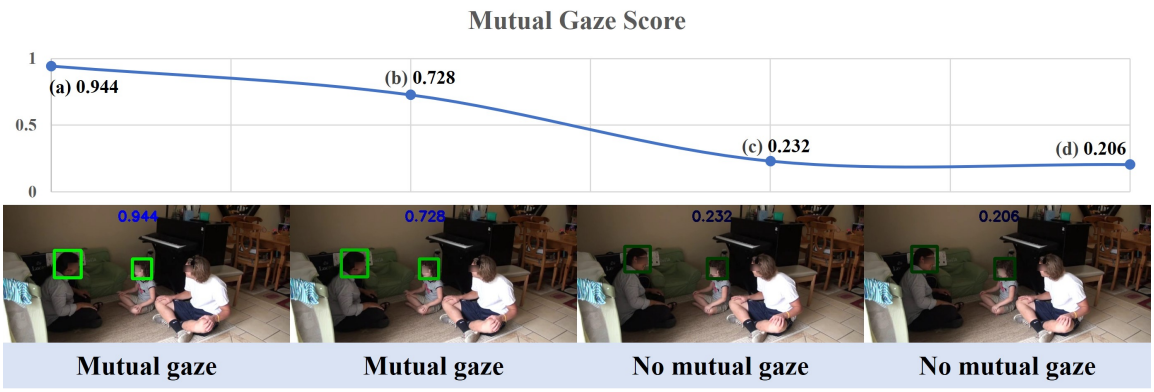


Figure 5.3: Mutual gaze detection results for child-trainer pair in four sample frames. When the child is looking down to the floor, the mutual gaze score is decreasing. We chose the cut-off point of 0.6 for mutual gaze features, since frames with values lower than this threshold do not show a mutual gaze feature.Mutual gaze score is in [0–1] range.

consists of 300 video clips with five different human interaction classes from 20 TV shows, and successfully detects mutual gaze features in different illuminations, scales, and cluttered background scenarios.

For our study in this chapter, we implemented the framework with Keras and TensorFlow on the AWS platform using one Tesla V100 GPU. The framework was pre-trained with AVA/UCO-LAEO dataset [5] including videos from 298 movies and 4 TV shows with annotated heads with bounding boxes, and AFLW dataset [206]) containing about $25k$ annotated faces in images. During the test process, we set the track length $K = 10$. This three-branch head tracking framework pre-trained with AVA/UCO-LAEO dataset could achieve *Average Precision (AP)* = 91.8% on the TVHID dataset and outperform other methods by 2–3% [5].

5.2.3 Measures

There is a fair amount of gaze interactions between the child and the trainers during the therapeutic interventions. The child in the therapy needs to identify the trainers' gaze behaviors, and perform and maintain the gaze direction towards the trainers accordingly while learning, playing, turn-taking, and verbal expressing during

the activities. Here we describe two main measures to evaluate the validness of our framework as an effective tool for assessing the social behavior of children with autism in the therapy sessions.

Mutual gaze ratio. The outcome of our deep learning framework provides a mutual gaze score for each central frame. Here, mutual gaze score is defined as the possibility that the child and the trainer are looking at each other in the video track, based on the central frame of the detected track. We define *mutual gaze ratio* for each child record based on the frequency of the mutual gaze between the child-trainer pairs over the entire therapy session. The mutual gaze detection results between the trainers were ignored. To calculate the mutual gaze ratio, our framework counts the frames (moments) when the child and trainers had eye contact, e.g., frames that the mutual gaze score is higher than our threshold (0.6), and the number of moments was divided by the number of total frames for the child record (see Figure 5.3). The threshold of 0.6 was decided empirically with test images to predict the mutual gaze score with the highest possibility. Since it is a normalized measure of mutual gaze interaction over the total frames, the range of this measure is from 0 (no mutual gaze at all) to 1 (mutual gaze all the time). Again this measure is automatically generated by our mutual gaze detection framework, and this will be compared with the hand-coded measures that we describe in the following.

Social visual behavior. While the measure of mutual gaze ratio is calculated by our deep learning framework automatically, we also have a measure of *social visual behavior* for children's looking patterns (looking at trainers/objects/self or looking away), which is hand-coded by the therapy experts. 100% inter-rater reliability agreement was established between two coders for the looking ratings ranging from [1–4]. Each child had one single social visual behavior score, and therapy experts looked at the non-verbal social behavior attentiveness of the children, e.g., mutual gaze behaviors during the clinician's single session interaction (45–60 mins) in the pre-test to score this measure on the scale of 1 (rare and brief looks towards trainers to receive feedback) to 4 (frequent and sustained social gaze towards trainers to receive feedback)

with 0.5 intervals. For the comparison with the results of our framework (i.e., mutual gaze ratio), we normalized the social visual behavior scores from [1–4] to [0–1] range.

5.3 Results

In the following, we report descriptive and inferential statistics results about our measures described in the previous section. We particularly focus on how well the output from our mutual gaze detection framework can be associated with the hand-coded evaluation measure, e.g., the correlation between the mutual gaze ratio and social visual behavior scores. In other words, we examine whether the framework output is a good indicator to represent the level of children participant's social behavior, which can potentially replace the hand-coded social behavior score by the therapy experts.

Comparison of descriptive stats. In order to investigate the effectiveness of our framework towards therapy interventions, we evaluate the social visual behaviors of the children in different therapy settings using both the mutual gaze score from our framework and the hand-coded score from the therapy experts. By simple mean comparison, the result shows that the mutual gaze ratio score was higher in the Play Therapy group ($M = 0.330$, $SD = 0.149$), compared to the Standard Therapy group ($M = 0.225$, $SD = 0.120$). This positive trend for the Play Therapy group aligns well with the result on the social visual behavior scores, which also shows a higher score for the Play Therapy group ($M = 0.500$, $SD = 0.279$) than the Standard Therapy group ($M = 0.417$, $SD = 0.419$). This suggests that our mutual gaze detection framework is effective for the provision of relevant information about the participants' social behavior. Table 5.2 provides the summary of these descriptive stats about the mutual gaze ratio and social visual behavior.

Correlation between mutual gaze ratio and social visual behavior. We further investigated the correlation between the mutual gaze ratio and social visual behavior measures. We used a non-parametric Spearman's rank correlation coefficient considering the ordinal type of raw social visual behavior and the small sample size. The result indicated a strong rank correlation between the mutual gaze ratio and

Table 5.2: Summary of mutual gaze ratio and social visual behavior scores in different therapy groups.

Group	# of Children (N)	Mutual gaze ratio $M \pm SD$	Social visual behavior $M \pm SD$
Standard Therapy	4	0.225 ± 0.120	0.417 ± 0.419
Play Therapy	6	0.330 ± 0.149	0.500 ± 0.279

Mutual gaze ratio is in [0–1] range.
Social visual behavior is normalized in [0–1] range.

the hand-coded social visual behavior score with statistical significance ($F_{(1,8)} = 2.53$, $p < 0.05$, $r_s = 0.650$) (see Figure 5.4). This again supports the mutual gaze ratio is a good indicator for the social visual behavior, which strengthens the effectiveness of our mutual gaze detection framework as an assessment tool for autism therapy.

Regression model for predicting social visual behavior from mutual gaze ratio. Finally, we tested how accurately we can predict the social visual behavior score (hand-coded by therapy experts) using the automated mutual gaze ratio from our framework and the participant profile information. Participant profile information was only collected through standard tests and surveys, which we described in Section 5.2. We particularly employed two profile data: *function* and *verbal* skill scores, which are directly related to the participants' cognitive ability and social behavior (see Table 5.1 for the details). We were interested in the benefit of our framework outcomes (mutual gaze ratio) as an important feature to predict the social visual behavior score; thus, we compared the prediction scores among four settings: (1) ground truth social visual behavior score, which the therapy experts hand-coded, (2) prediction based on our model with the mutual gaze ratio from the framework outcome together with the function and verbal profile scores, (3) prediction only based on those two profile scores, and (4) random prediction.

A machine learning method was applied for regression-based prediction. We selected a basic linear regression model using Mean Squared Error (MSE) as the loss function to measure the difference between predicted and actual social visual behavior scores. To make up for the small data size, we adopted the Bootstrap approach to

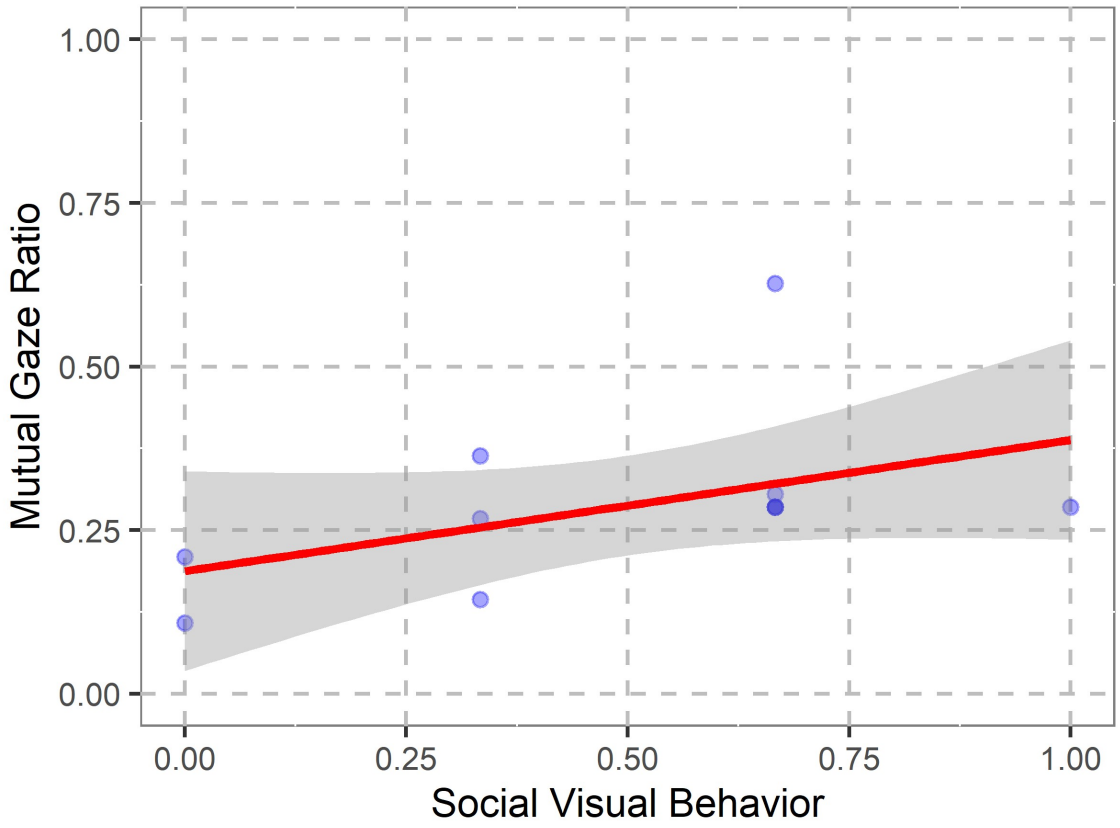


Figure 5.4: The scatter plot of mutual gaze ratio with hand-coded social visual behavior score from ten children. Mutual gaze ratio is in [0–1] range. The social visual behavior score is normalized in [0–1] range. The point in a darker color indicates two points overlapped. The positive trend with these two features is present in this plot.

quantitatively control and check the stability of the results by random sampling with replacement. The re-sampled dataset was used for training by leave-one-out cross-validation with a learning rate at 0.01. The results were promising to support the benefit of our model using mutual gaze ratio outcome with the profile information for better prediction of the social visual behavior, compared to the other two models with the (3) and (4) settings. Figure 5.5 shows the ground truth and the prediction values of our regression model and two other comparison models. According to the bar chart, the results of our regression model were closer to the ground truth values than the results of the other models with function and verbal scores and the random values. Our model showed a lower MSE-loss of 0.177 compared to the model only

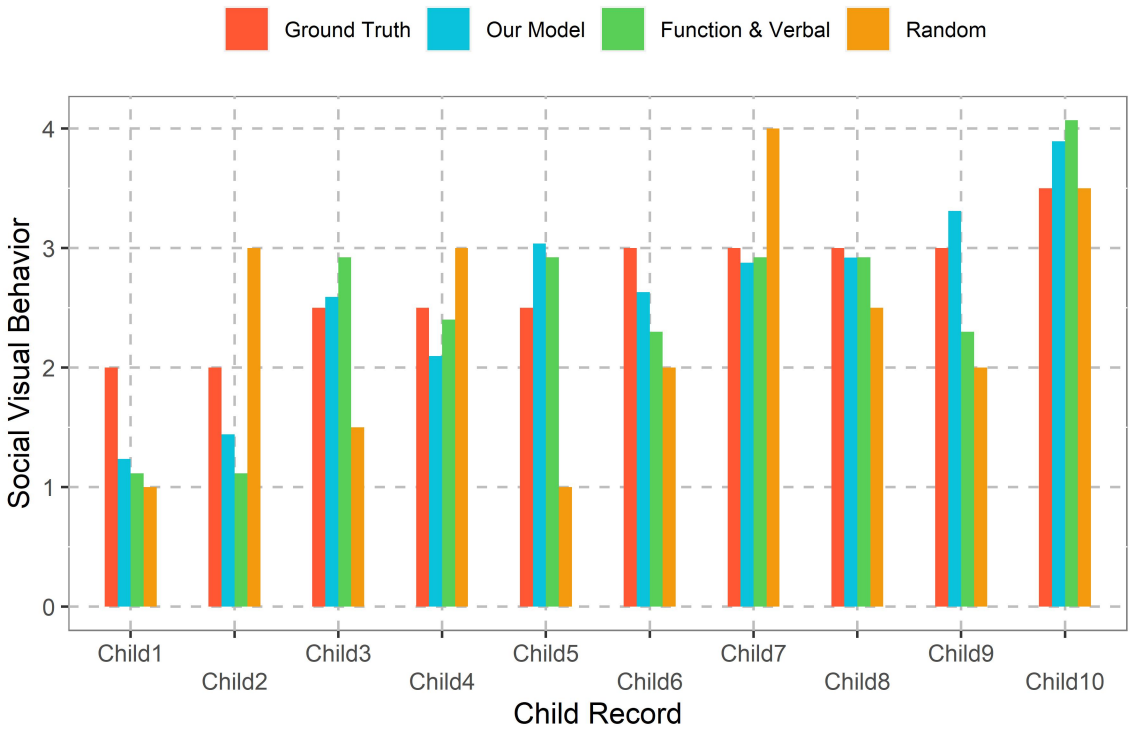


Figure 5.5: The bar chart of the predicted social visual behavior scores (in [1–4] range) over the child records: ground truth hand-coded by therapy subject-matter experts (red); prediction based on participants profile on function and verbal skills (green); prediction based on our work, that uses the mutual gaze ratio and participant profiles (blue); and random prediction (orange). The blue bars (our model) resemble ground truth points of social visual behavior more accurately than the rest.

based on function and verbal skill information ($MSE\text{-}loss = 0.315$) or the random prediction ($MSE\text{-}loss = 0.850$). The weight corresponding to the mutual gaze ratio feature ($W_{mg} = 2.046$) was also higher than the weights for the other two profile features ($W_f = 0.748$, $W_v = 0.266$), which suggests that mutual gaze features provide a more positive impact for the prediction than the function and verbal levels. The positive value of the weight also represents a positive relationship between the mutual gaze ratio and the predicted value—social visual behavior score. This indicates the outcome of our mutual gaze detection framework can be an important feature to predict the social behavior score.

5.4 Discussion

Capturing children's gaze interaction during the therapeutic training activities, particularly about the mutual gaze towards the trainers, can reveal valuable social visual information; thus, analyzing such behaviors is highly useful for effective therapy treatments and evaluations. While there have not been many studies that adopted deep learning-based approaches in video-based autism intervention contexts, we here proposed a mutual gaze detection framework using the three-branch head tracking approach for autism therapy, based on a state-of-the-art gaze detection method [5]. To evaluate our proposed framework, we conducted a series of analyses that compared the outcomes of the framework with manually-coded ground truth measure of social visual behavior, with an in-house ASD dataset collected from the autism therapy intervention [15].

As we reported in the previous section, our results showed that our framework could successfully provide mutual gaze ratio scores, which were comparative to the social visual behavior scores. For different therapy settings, descriptive comparison results of our method were matching the results of hand-coded measures (social visual behavior score), which were annotated by the therapy experts. We also found a significant result, which shows a strong positive rank correlation between mutual gaze ratio and social visual behavior score. Moreover, our framework was able to predict the social visual behavior score using the mutual gaze ratio and other participant-profile features with a simple linear regression model. This means that our method can be used as a novel assessment tool for automatic and reliable mutual gaze detection for social visual behavior evaluation in autism therapy interventions. Since the social gaze behaviors are scored based on various social gaze patterns, to reduce the loss, more social gaze patterns need to be collected, measured, and used for predictions. Compared with typical time-consuming hand-coding approaches to examine the visual performance of children with autism, our method was able to effectively and efficiently extract mutual gaze features during the whole training process with a standard camera.

However, there are still certain limitations in our work as well, which can help

us expand on this research in the future. This work was still a preliminary study with a limited number of samples. Like most autism studies, recruitment of children with autism was not easy, and that made the balancing of samples across groups difficult. There is still a large subset of collected data that we need to include in our future work, which was excluded from the current study due to the absence of social visual behavior annotations. Also, the framework [5] we adopted in this chapter detected the mutual gaze features based on head tracks, which may not be able to capture the correct gaze information for specific situations, e.g., in the case that the eye gaze is not matched with the head direction or for people with eye movement disorders. Detected accuracy may also be affected by camera occlusion and video low resolution. We plan to investigate and develop more accurate eye gaze detection methods to overcome this issue. Finally, our method has been evaluated only at a system level. Further evaluation and discussion with subject-matter experts, in this case autism therapists, would be invaluable to design and develop better gaze detection or other social behavior evaluation tools in practice.

5.5 Conclusion

In this chapter, we introduced a deep learning mutual gaze detection framework using our in-house ASD dataset, as an effective tool for automatic prediction of social visual behavior scores of children with autism in the context of play therapy. The effectiveness of the framework was validated by comparing the outcome with social visual behavior scores hand-coded by therapy experts. Our results using different analysis approaches, including descriptive comparisons, correlation analysis, and regression predictions, showed that our proposed framework was able to automatically extract mutual gaze features during the therapy sessions, and could reliably represent, even accurately predict social behavior scores, comparable to human experts.

Our findings have implications in special education technology and autism therapy analysis by offering a novel assessment tool for social visual behaviors based on gaze information. Beyond the play therapy context, our framework can be applied to other

application scenarios that need reliable automatic social behavior analysis from videos involving human-human or human-virtual agent interactions [16, 207, 208]. Also, we aim to extend our framework via semi-supervised and unsupervised machine learning methods to analyze the non-annotated video recordings. Moreover, we plan to extend the behavioral features of our framework, such as gaze sharing and following [39, 47], facial expressions [209, 153], and body gestures [210, 211] for more comprehensive interpretation of social visual behaviors, including social engagement and social anxiety [212] of children in the spectrum.

Chapter 6

AUTISM SOCIAL VISUAL BEHAVIOR ANALYTICS BASED ON ADVANCED MUTUAL GAZE DETECTION

6.1 Problem Statement

As a type of non-verbal communication, social visual behavior plays a central role in studying social cognitive processes in interactive and complex settings. Visual cues are mostly displayed and perceived spontaneously by humans [213], and it is commonly believed that the level of cognitive abilities and social skills are reflected in gaze behavior [214]. More specifically, in special education, individuals and children with Autism Spectrum Disorder (ASD) have difficulty in identifying, performing, and maintaining such gaze behaviors [30]. The lack of such social engagement may lead to anxiety, depression, and social avoidance in individuals on the Autism spectrum during social interactions [31]. Individuals on the autism spectrum have poor quality of social interactions, e.g., reduced eye contact [32], reduced interest in social stimuli [33], lack of response to name [34], and insufficient sharing of interests [35]), and lack of social connections with partners [36]. Mutual gaze, or two people looking at each other, is an important type of social gaze [32] and has been considered a critical indicator for establishing and maintaining successful face-to-face interactions during everyday interactions [38]. In many autism therapy interventions, mutual gaze has been used as an objective measure to interpret the social behaviors of children with autism and evaluate therapy effectiveness [40, 41].

In Chapter 5 we investigated the linkage between mutual gaze and social visual behavior. However the tracking method had limited accuracy and we didn't do further social behavior analysis since autism dataset was relatively small and subjective to bias

due to imbalanced groups. In this chapter, we introduce a social visual behavior analytics approach for observing the mutual gaze performance of children with autism during the therapy sessions using mutual gaze detection. This work uses the expanded balanced data with more representative samples collected from an in-house play therapy intervention for children with autism ($N = 21$ children, 7.57 ± 2.31 of years old). The training performance of each child was captured by the standard camera and recorded as video clips. To automatically extracted the mutual gaze features from the video-based data, We further enhanced the gaze detection model using an advanced mutual gaze detection framework [6] and generated the social visual behavior measures based on the framework outcomes. The effectiveness of our framework was demonstrated by comparing the mutual gaze ratio generated based on the framework outcomes with human-coded ratio measures annotated by the therapy experts. The effectiveness of our framework was assessed by comparing the mutual gaze ratio generated based on the framework outcomes with hand-coded ratio measures annotated by human experts. The social visual behavior of children with autism was examined across different therapy settings, training activities, and therapy sessions. We also predicted the social visual behavior score using multiple machine learning-based regression models with on our mutual gaze-related measures. We integrated more ASD-related measurements from therapists, such as AODS-2 social affect score, into our computational models. We found that the random forest model achieved the best performance, assisted by the level of functioning skills and social affect score from the Autism Diagnostic Observation Schedule (ADOS [215]).

The remainder of this chapter is structured as follows. Section 6.2 as the materials and method section describes the play therapy intervention and our in-house ASD dataset, deep learning methodology for mutual gaze detection, and measures for further social behavior analytics. Section 6.3 presents our results and findings, that is discussed in Section 6.4. Section 6.5 concludes the chapter and informs future research directions and research implications.

6.2 Materials and Method

In this section, we describe the details of our deep learning-based mutual gaze detection framework for gaze behavior extraction, provide the extended in-house ASD dataset we collected, and introduce our measures, which we used for social visual behavior analysis.

6.2.1 Autism Therapy Interventions and Expanded In-House Data Collection

In order to develop a social visual behavior analysis approach, particularly for autism therapy efficacy, we collected a number of video clips of different autism therapy training interventions through a multi-session human subjects study [15]. Compared with the study introduced in Chapter 5, our study in this section clearly introduced three different activities within two group settings with more children enrolled. The therapy intervention enrolled a total of 36 children with autism in a randomized controlled trial study. The study was approved by the university's Institutional Review Board # 637082-12. Children participants were recruited through flyers posted online and onsite in local schools, services, and self/parent advocacy groups, and were randomly assigned to one of the therapy groups.

There totaled ten weeks of intervention, with the pre-test and post-test sessions conducted in the first and the last weeks, and the autism therapy training activities provided for two therapy sessions per week during the intermediate eight weeks. In the first week, children completed pre-test sessions, and the therapists determined their social, communication, and functional abilities based on their pre-test performance. During the following eight weeks, they had two therapy sessions per week. For therapy sessions, two trainers interacted with the child to complete embodied creative activities in the child's home environment setting. One trainer introduced the details of the activity and provided guidance, and the other trainer practiced the activity with the child as a model. All the trainers, as pediatric physical therapists or graduate students/faculties, were well trained by autism therapists. With parents' permission and

the notification to the children, the therapy sessions were videotaped with a standard camera, which was located at a fixed position toward the child. Finally, the post-test session was completed by all children in the last week to examine changes in social, communication, and functional abilities.

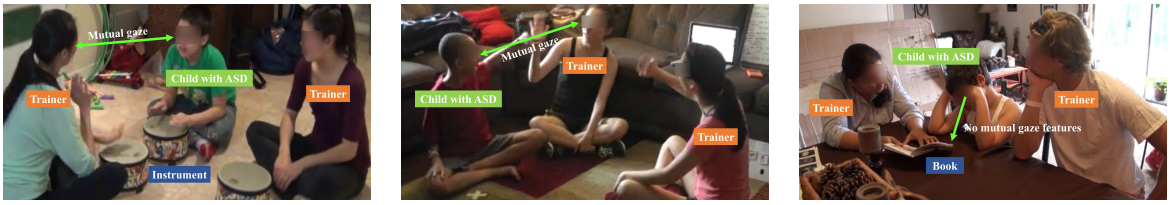
In this chapter, we particularly focus on children participants' visual interaction with other people during the training activity. Therefore, two therapy groups with 28 children were selected as a specific subset of video data.

- **Play Therapy**: In the Play Therapy group, children participants were engaged in an improvisational "Music Making" activity using musical instruments, like drums, xylophones, cymbals, tambourines, etc. (see Figure 6.1a as an example) and a "Hello Song" activity involving singing hello or action songs (see Figure 6.1b as an example). During the therapy session, the child was guided to follow the instructions, copy the trainers' movements, play the instruments with the musical beats, make music with the trainers, or sing a hello/action song along with hand clapping and waving gestures. The themes of this session included various rhythms and musical components, such as start and stop, steady beat, turn-taking, slow and fast, soft and loud, and collaborative music creation.

- **Standard Therapy**: In the Standard Therapy group, children participants were engaged in a table-top reading activity using age-appropriate books (see Figure 6.1c as an example). During the therapy session, the child was guided to follow the instructions, take turns with the trainers to read books, answer book-related questions, as well as spontaneous expressions. Session themes included social communication and reading books about people, their things, food, transportation, etc.

In the Play Therapy group, one trainer was seated opposite to the child and the buddy sat next to the child, while in the Standard Therapy group, both trainers were seated next to the child around the table. While both groups engaged in social communication activities such as eye contact, turn-taking, and non-verbal/verbal communication, the Standard Therapy group mainly focused on table-top reading, whereas the Play Therapy group also focused on gross motor skills. [63, 64].

Expanded in-house ASD dataset. We prepared an expanded in-house ASD dataset for our social visual behavior analytics based on a subset of the collected video data from the autism therapy study described above. The data were collected



(a) (b) (c)

Figure 6.1: Sample scenes captured from two therapy groups in our ASD dataset: Example frames of the Play therapy group during (a) the "Music Making" activity and (b) the "Hello Song" activity with mutual gaze behavior; example frames of the Standard Therapy group during (c) the "Reading" activity without mutual gaze behavior. Our study in this chapter aims to automatically capture mutual gaze attentiveness of children with autism, and analysis their social visual behavior.

in practical in-the-wild settings where therapists conducted sessions without knowing our analysis method and mutual gaze detection framework, which made the balanced samples difficult. For our dataset, 7 out of 28 children's video records were excluded due to low resolution, camera occlusion, or incomplete data records or video annotations. Each record in our data included the video clips from sessions 1, 8, and 16, which covered the start, middle, and last sessions of the therapy. Other therapy sessions (13 video clips for each child record) were excluded from the analysis due to the lack of annotation. After post-processing, our ASD dataset had a total of 21 hours of video from children in 84 video clips (30 for the Standard Therapy group with "Reading" activity, 33 for the Play Therapy group with "Music Making" activity, and 21 for the Play Therapy group with "Hello Song" activity) with 21 children (Standard Therapy: 10, Play Therapy: 11). Seven children from the Play Therapy group did both the "Music Making" activity and the "Hello Song" activity. Each video clip was 25 fps with approximately 15-minute length. We used the ASD dataset as the input of our mutual gaze detection framework to calculate the children's mutual gaze score during the therapy session for further visual behavior analyses.

Participants profile surveys. Before each child's enrollment, parental consent, the Social Communication Questionnaire (SCQ [202]), and the ADOS-2 [215] were completed to confirm their eligibility. The survey data included the demographic

Table 6.1: Demographic characteristics of children in our ASD dataset.

Participant Characteristics	Play Therapy	Standard Therapy	F or χ^2 value	p-value
Age ($M \pm SD$)	7.82 ± 2.52	7.30 ± 2.16	0.88	0.55
Gender	8 M, 3 F	10 M, 0 F	3.18	0.07
ADOS-2 social affect score ($M \pm SD$)	17.27 ± 4.56	16.30 ± 5.64	1.88	0.20
Level of Functioning ($M \pm SD$)	1.91 ± 0.83	2.30 ± 0.82	0.58	0.57

Level of functioning skill is scored in range from 1 (low) to 3 (high).
Except for gender, the demographic information of children in our dataset is balanced for two therapy groups.

information of children with autism, including age, gender, ADOS-2 social affect score, and level of functioning measures. The age range of the 21 children (3 females) included in our ASD dataset was between 5 and 12 years. The level of functioning of each child was based on their level of independence during daily living skill and was scored by the therapists on a range from 1 to 3 (1: low functioning or or needing significant support, 2: medium functioning needing moderate support, and 3: high functioning or needing less support). See Table 6.1 for demographic characteristics of children in our ASD dataset. Except for gender, the demographic information of children in our dataset is balanced for two therapy groups.

6.2.2 Advanced Mutual Gaze Detection Framework

Uisng a video recordings of therapy sessions, we aim to develop a deep learning-based framework that can automatically determine if there is any mutual gaze type of social visual interaction between the therapy trainers and the children participants. The framework would be practical and computational for the therapists and trainers to observe the state of the children participants during the entire training process. To this end, we describe the development of our deep learning-based mutual gaze detection framework, which adopted a state-of-the-art three branch track network [6] on our in-house ASD dataset to extract mutual gaze features between all possible pairs of people from the video clips.

Unlike most methods that rely on faces for gaze estimation, considering there

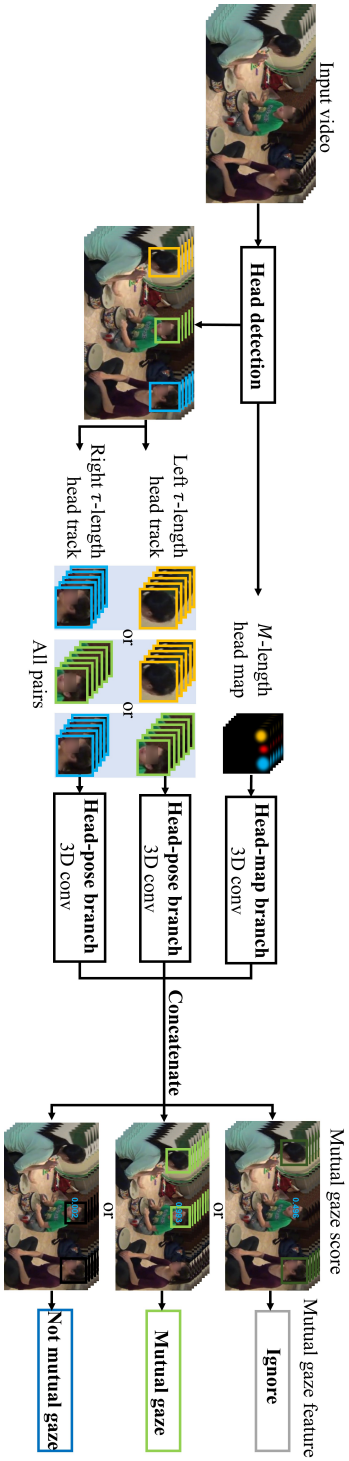


Figure 6.2: Our deep learning architecture for mutual gaze detection, adopted from the three-branch head tracking framework [6] on therapy videos from the ASD dataset. Using mutual gaze scores from the framework outcome, we can recognize mutual gaze features between the child-trainer pairs (trainer-trainer pairs are ignored). The light green bounding boxes are shown with a high mutual gaze score; the dark green/black bounding boxes are shown with a low mutual gaze score, which represents no mutual gaze feature is detected from the scene.

is no guarantee that the eyes are fully visible in the collected scenes, our framework focuses on detecting the head pose (i.e., position and orientation) of all the people from the frames. Firstly, the input video clip is sent to the Single Shot Multibox Detector (SSD [204]) for head detection. Then the linking algorithm [203] groups them into tracks as the input of the three branch track network. The network has three branches, including two head-pose branches and one head-map branch (see Figure 6.2). Each head-pose branch encodes a tensor of τ RGB frame crops of size 64×64 pixels, which contains the head sequence of one person of the target pair, taking into account the head pose. The head-map branch embeds relative head positions and relative distance to the camera (i.e., depth) between two head tracks over time using a 64×64×M map with 2D Gaussians for the whole τ-frame track. This framework extends the temporal dimension of the head maps and considers multiple consecutive frames instead of single frames in order to reduce the influence of noise, inconsistency, and detection problems. The different Gaussian sizes of the head map are proportional to the head sizes (i.e., detected bounding boxes), which encode the relative 3D arrangement (depth) of the people in the scene. In addition to the two head tracks, the position information for other persons in the scene is also encoded by the head-map branch in order to detect the case where the third person cuts the gaze ray between the two-side people. Each head-pose branch consists of five 3D convolutional layers, and the head-map branch consists of four 3D convolutional layers instead of the 2D ones in other works. After applying L2-normalization, the outcome embedding vectors from different branches are concatenated and further processed by a fusion block, which includes a fully connected layer with a dropout layer and a Softmax layer to output a confidence score on whether the target pair of people have eye contact or not, i.e., the score for the mutual gaze. This framework applies to all pairs of simultaneous head tracks in the video clips.

The outcome of the framework consists of all frames with detected mutual gaze scores for all possible pairs. In each frame of the track, there are bounding boxes for the heads of each pair with the light or dark green color based on the confidence score (see Figure 6.3, for example). The score shows how likely the pair of the target pair

is to have eye contact interaction. In other words, the higher the score is (the lighter the color of the bounding boxes is), the higher likelihood the people are looking at each other. According to the study conducted in the original paper [6], this framework achieves the state-of-the-art results on the TVHID dataset [205], which consists of 300 video clips with five different human interaction classes from 20 TV shows, and successfully detects mutual gaze features in different illuminations, scales, and cluttered background scenarios.

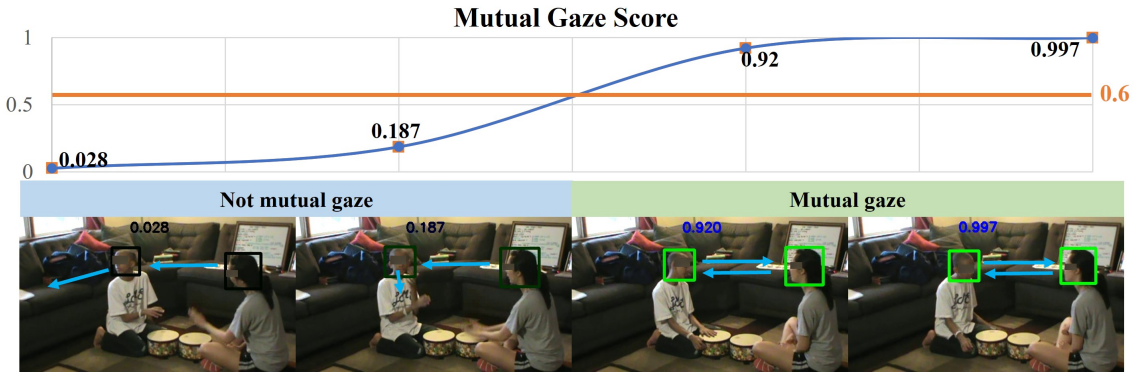


Figure 6.3: Mutual gaze detection results for child-trainer pair in four sample frames. When the child is looking toward the trainer, the mutual gaze score is increasing. We chose the cut-off point of 0.6 for mutual gaze features, since frames with values lower than this threshold do not show a mutual gaze feature. Mutual gaze score is in [0–1] range.

In this work, we implemented the framework with Keras and TensorFlow on the AWS platform using one Tesla V100 GPU. The framework was pre-trained with AVA/UCO-LAEO dataset [6] including videos from 298 movies and 4 TV shows with annotated heads with bounding boxes, and AFLW dataset [206]) containing about $25k$ annotated faces in images. During the test process, we set the head track length $\tau = 10$ and head map length $M = 10$. This framework pre-trained with AVA/UCO-LAEO dataset could achieve *Average Precision (AP)* $= 92.3\%$ ($M = 1$) for the TVHID dataset and outperform the previous framework by 4–7% for UCO-LAEO and around 18% for AVA-LAEO [6]. Since our study in this chapter particularly focuses on the social visual behavior of children participants, the mutual gaze detection results between the

trainers were ignored. The framework automatically detected and generated mutual gaze scores in 10–20 minutes for each video clip, which saves more time and effort than human annotation (1–2 hours for each video clip according to our subject matter expert's input).

6.2.3 Measures

In the therapy sessions, children looked in the direction of trainers, and maintained their gaze towards the trainers while learning, playing, turn-taking, collaborating, and speaking to trainers during the activities. To analyze the social visual behavior of children participants in the therapeutic interventions, gaze interactions between the child and the trainers can be detected by our mutual gaze detection framework. Here we describe the main measures for framework evaluation and social visual behavior analytics. In addition to the measures used in Chapter 5, we defined mutual gaze duration and human-coded ratio as new measures.

Mutual gaze ratio. The outcome of our mutual gaze detection framework provides a mutual gaze score for each frame. The mutual gaze score of child-trainer pairs is defined as the possibility that they are looking at each other in the frame based on the detection of the track. Here, we define *mutual gaze ratio* for each child based on the frequency of the mutual gaze between the child-trainer pairs over the therapy sessions. To calculate the mutual gaze ratio, our framework counts the frames (moments) when the child and trainers had eye contact (in the session 1, 8, and 16), e.g., frames that the mutual gaze score is higher than our threshold (0.6), and the number of moments was divided by the number of total frames for each child (see Figure 6.3). Since it is a normalized measure of mutual gaze interaction over the total frames, the range of this measure is from 0 (no mutual gaze at all) to 1 (mutual gaze all the time).

Mutual gaze duration. Based on the mutual gaze scores provided by the deep learning-based mutual gaze detection framework outcome, we define *mutual gaze duration* for each child based on the average duration of the mutual gaze between

the child-trainer pairs throughout therapy sessions. Our framework counts the frames (moments) when the child and trainers maintained eye contact for more than one second (in the session 1, 8, and 16), e.g., the consecutive frames (more than 25 frames) that the mutual gaze score is higher than our threshold (0.6), and compute the mean value of the consecutive frame numbers for each child.

Human-coded ratio. While the measure of mutual gaze ratio is calculated by our deep learning framework automatically, we also have a measure of *human-coded ratio* for the frequency of children's social gaze performance during the session 1, 8, and 16, which is hand-coded by the therapy experts as the ground truth. Since the trainers performed interactively all the time and their gaze directions were toward the child for most of the time during the therapy sessions, the hand-coded mutual gaze annotation only focused on the children's social gaze behaviors. The human-coded ratio is calculated by the time (second) the child looking at the trainers divided by the total time (second) of entire therapy sessions. As a normalized measure of social gaze interaction over entire sessions, the range of this measure is from 0 (never looking at the trainers) to 1 (looking at the trainers all the time). This measure, as the ground truth, is compared with the mutual gaze ratio generated from the mutual gaze detection framework outcomes for framework evaluation.

Social visual behavior score. *Social visual behavior score*, as a measure of for children's looking patterns (looking at trainers/objects/self or looking away), is hand-coded by the therapy experts. Each child had one single social visual behavior score, and therapy experts looked at the non-verbal social behavior attentiveness of the children during the clinician's single session interaction (45–60 mins) in the pre-test to score this measure on the scale of 1 (looking away and lack of gaze interaction with trainers) to 4 (frequent and sustained social gaze towards trainers to receive feedback) with 0.5 intervals.

100% inter-rater reliability agreement was established between two coders for the participants' looking patterns (looking at trainers/objects/self or looking away) based on 20% of video data. Both human-coded mutual gaze ratio and social visual

behavior score are derived from the annotated looking patterns.

6.3 Results

In the following, we report descriptive and inferential statistics results about our measures described in the previous section. We evaluate our framework by comparing the mutual gaze ratio automatically generated from the mutual gaze detection framework outcomes with the human-coded ratio annotated by the human experts. Then, we analyze children's social gaze behavior across different therapy groups, training activities, and therapy sessions. Finally, we predict the social visual behavior score using different machine learning-based regression models. In our results, significance was set at $p \leq 0.05$.

Framework evaluation. We first report descriptive and inferential statistics results about our measures described in the previous section. In order to investigate the effectiveness of our framework towards therapy interventions, we evaluate the social visual behaviors of the children in different therapy settings and activities using both the mutual gaze ratio from our framework and the human-coded ratio from the human experts (see Table 6.2). By mean comparison, the results show that the mutual gaze ratio was higher in the Play Therapy group ($M = 0.118$, $SD = 0.079$), compared to the Standard Therapy group ($M = 0.101$, $SD = 0.058$). This positive trend for the Play Therapy group aligns well with the result on the human-coded ratio, which also shows a higher ratio for the Play Therapy group ($M = 0.285$, $SD = 0.171$) than the Standard Therapy group ($M = 0.193$, $SD = 0.162$). However, the difference between the two group settings is not significant for mutual gaze ratio ($p = 0.539, t_{(26)} = -0.622, ns$—$ns$ stands for statistically non-significant) or human-coded ratio ($p = 0.182$, $t_{(26)} = -1.374, ns$) based on the independent t-tests.

We further investigated the association between the mutual gaze ratio and human-coded ratio measures by calculating the Pearson correlation coefficient [216]. The result indicated a strong positive linear relationship between the mutual gaze ratio and the human-coded ratio with statistical significance ($F_{(1,26)} = 18.33$, $r_p = 0.643$,

Table 6.2: Summary of mutual gaze ratio, duration, and human-coded ratio in different therapy groups and activities.

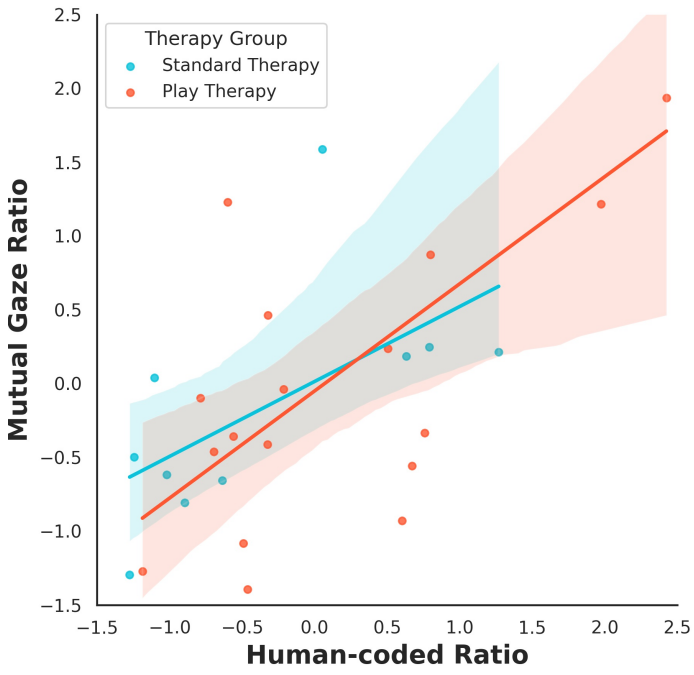
Group	Activity	# of Observation (N)	Mutual Gaze Duration ($M \pm SD$)	Mutual Gaze Ratio ($M \pm SD$)	Human-coded Ratio ($M \pm SD$)
Play Therapy	Hello Song	7	53.40 ± 12.02	0.166 ± 0.095	0.403 ± 0.191
	Music Making	11	57.44 ± 10.44	0.088 ± 0.052	0.210 ± 0.109
	Combined	18	55.82 ± 10.87	0.118 ± 0.079	0.285 ± 0.171
Standard Therapy	Reading	10	61.07 ± 11.45	0.101 ± 0.058	0.193 ± 0.162
Total		28	57.92 ± 11.18	0.112 ± 0.072	0.252 ± 0.171

Mutual gaze ratio, as the frequency of the mutual gaze between the child-trainer pairs over the therapy sessions based on the detected mutual gaze scores and the total frames of the task, is in [0–1] range.
Human-coded ratio, as the frequency of the mutual gaze between the child-trainer pairs over the therapy sessions based on hand annotations and the total time (second) spent on task, is in [0–1] range.
Mutual gaze duration is the mean value of the frame numbers when the child is maintaining a social gaze behavior over 25 frames (1 second).

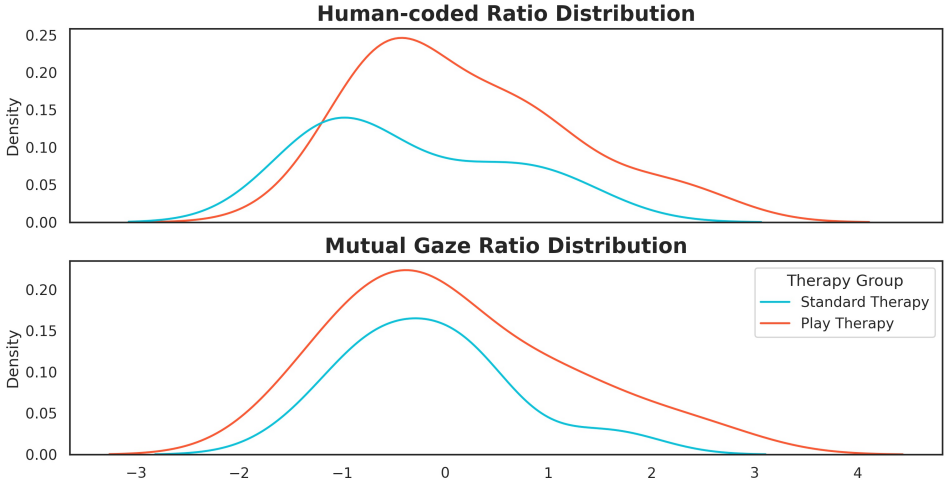
$p < 0.0005$, $RMSE = 0.056$). The linear relationship and distributions of the standardized mutual gaze ratio and human-coded ratio are shown in Figure 6.4. As the similarity of these two distribution highlights, we can suggest that mutual gaze ratio is a reliable indicator of the social visual behavior. Thus, it shows the effectiveness of our mutual gaze detection framework as an automated assessment tool for evaluating and representing the social behavior of autism.

Social visual behavior analytics for children with autism. We analyzed the social visual behavior of the children across therapy settings (Play Therapy and Standard Therapy), within-group activities ("Music Making" activity and "Hello Song" Activity in the Play Therapy group), and therapy sessions (early and late sessions) using the measures described in section 6.2.

Independent t-tests were used to compare the mutual gaze performance (mutual gaze ratio and mutual gaze duration) of children between different therapy groups. Although the children in the Play Therapy group got higher mutual gaze ratio ($M = 0.118$, $SD = 0.079$) than the Standard Therapy group ($M = 0.193$, $SD = 0.162$), there were no significant between-group difference on the mutual gaze ratio ($p = 0.539$,$t_{(26)} = -0.622$, ns) , and neither on the mutual gaze duration ($p = 0.259$, $t_{(23)} = 1.157$, ns).



(a)



(b)

Figure 6.4: (a) The scatter plot of mutual gaze ratio and human-coded ratio from 28 observations in different therapy group settings. The regression line and 95% confidence interval (shaped area) for each group are also included. (b) The distributions of mutual gaze ratio and human-coded ratio in different therapy group settings using kernel density estimation. The distributions of two ratios are very similar for both therapy group settings. Both ratios are standardized.

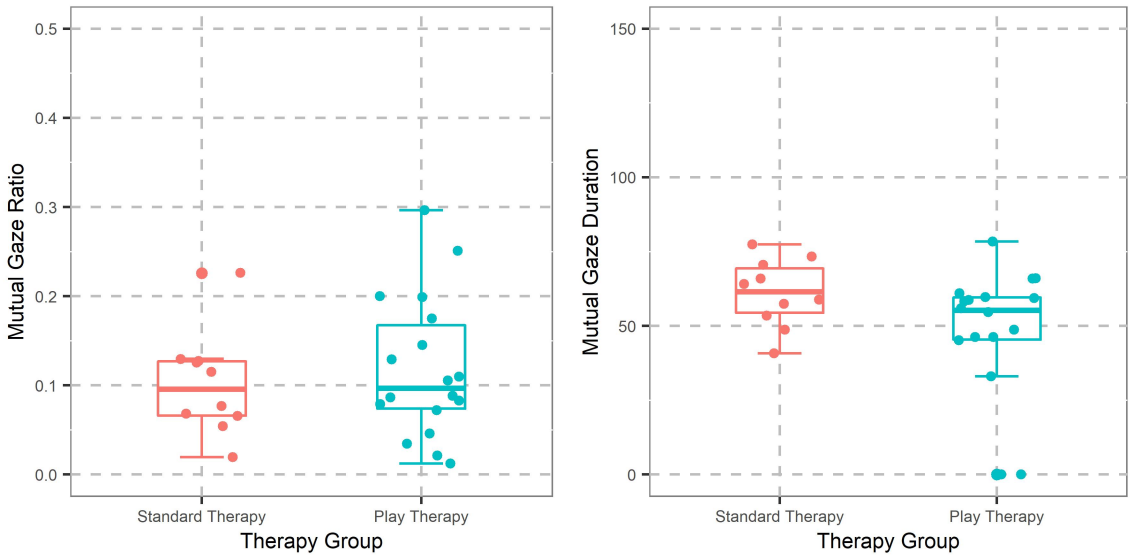


Figure 6.5: The box plots of mutual gaze ratio and mutual gaze duration in the Play Therapy group and the Standard Therapy group. No significant between-group difference on the mutual gaze ratio or duration. Mutual gaze ratio is in [0–1] range.

Figure 6.5 shows the box plots of mutual gaze ratio and mutual gaze duration in different group settings.

To investigate the effects of within-group activities, independent t-tests were used to compare the children's mutual gaze performance between the "Music Making" activity and the "Hello Song" activity in the Play Therapy group. There was a statistically significant within-group difference on the mutual gaze ratio ($p < 0.05$, $t_{(16)} = -2.28$, $Hedge's\ g = -1.05$) between the "Hello Song" activity ($M = 0.166$, $SD = 0.095$) and the "Music Making" activity ($M = 0.088$, $SD = 0.052$) with a large effect size (see Figure 6.6). On the contrary, no significant within-group differences was found on the mutual gaze duration ($p = 0.501$, $t_{(13)} = 0.693$,ns) between the "Hello Song" activity ($M = 53.40$, $SD = 12.02$) and the "Music Making" activity ($M = 57.44$, $SD = 10.44$). The results showed that different type of activities provided varieties of within-group effects on the frequency of the children's social gaze behaviors. Children performed gaze interaction with the trainers more frequently in the "Hello Song" activity than in the "Music Making" activity.

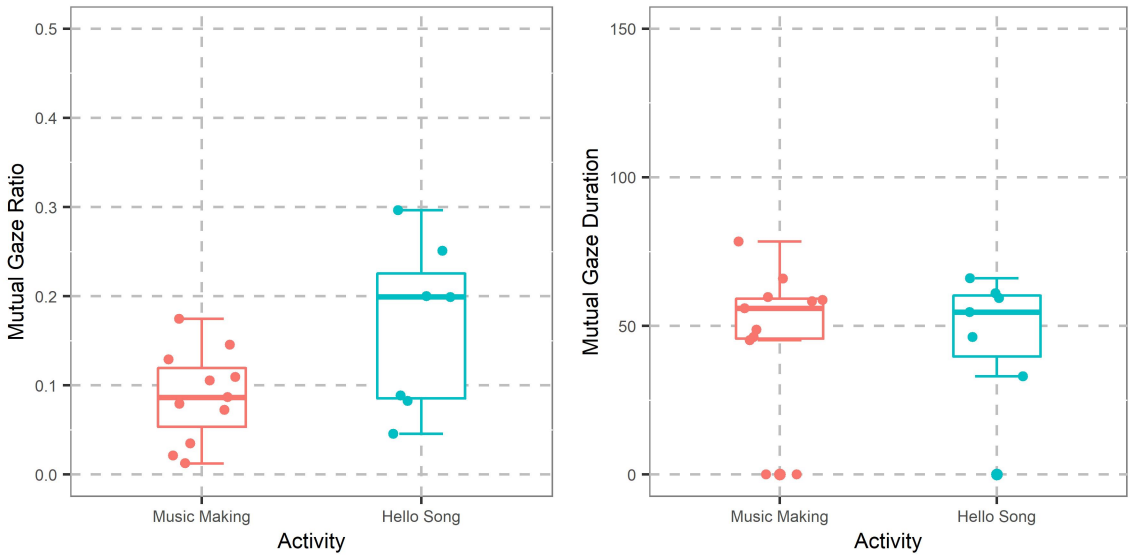


Figure 6.6: The box plots of mutual gaze ratio and mutual gaze duration in the "Hello Song" and the "Music Making" activities within the Play Therapy group. The mutual gaze ratio in the "Hello Song" activity is significantly higher than the ratio in the "Music Making" activity. No significant within-group difference on mutual gaze duration. Mutual gaze ratio is in [0–1] range.

Descriptive and inferential statistics results about mutual gaze measures in different therapy groups and activities across therapy sessions were reported in Table 6.3 and Figure 6.7. Independent t-tests were used to compare the mutual gaze performance (mutual gaze ratio and mutual gaze duration) of children across the early and late sessions (session 1 and 16). Results showed that neither mutual gaze ratio ($p = 0.928$, $t_{(54)} = -0.090$,ns) nor social visual duration ($p = 0.523$, $t_{(37)} = -0.645$,ns) had significant improvement after the therapy training process. However, compared to the mutual gaze performance in the early session, 1 observation in the Hello Song activity, 4 observations in the Music Making activity, and 5 observations in the Reading activity had higher mutual gaze ratios and longer duration in the late session.

Social visual behavior score prediction using machine learning models. We evaluated the performance of the social visual behavior score (manually scored by human experts) prediction based on our mutual gaze-related measures (mutual gaze

Table 6.3: Summary of mutual gaze ratio and duration in different therapy groups and activities across therapy sessions.

		Play Therapy		Standard Therapy	Total
		Music Making	Hello Song	Reading	
# of Observation (N)	Early Session	11	7	10	28
	Late Session	11	7	10	28
Mutual Gaze Ratio ($M \pm SD$)	Early Session	0.100 ± 0.102	0.227 ± 0.130	0.111 ± 0.134	0.136 ± 0.128
	Late Session	0.100 ± 0.137	0.191 ± 0.129	0.144 ± 0.098	0.139 ± 0.123
Mutual Gaze Duration ($M \pm SD$)	Early Session	50.26 ± 19.31	54.32 ± 11.69	57.74 ± 26.86	54.61 ± 22.04
	Late Session	57.46 ± 13.28	58.89 ± 29.02	59.28 ± 12.64	58.62 ± 16.15

Early and late sessions represent the therapy training session 1 and 16.
Mutual gaze ratio is in [0–1] range.
Mutual gaze duration is the mean value of the frame numbers when the child is maintaining a social gaze behavior over 1 second (25 frames).
Neither mutual gaze ratio nor duration had significant improvement after the therapy training process.

ratio and mutual gaze duration) and the participant profile information. We particularly employed two profile data: level of functioning and ADOS-2 social affect score, which are directly related to the children's cognitive ability and diagnosis of ASD (see Table 6.1 for the details). To investigate the benefit of our mutual gaze ratio measures as critical features to predict the social behavior score, we conducted an ablation experiment by comparing the prediction performance between two settings: (1) prediction based on our model with the mutual gaze ratio and duration from the framework together with the level of functioning and ADOS-2 social affect score, and (2) prediction only based on those two profile scores, using the following five regression models: random forest (RF), support vector regression (SVR), Lasso regression (Lasso), gradient boosting trees regression (GBT), and multi-layer perceptron regression (MLP). Mutual gaze ratio, mutual gaze duration, level of functioning, and ADOS-2 social affect score are normalized. To mitigate the training issues due to small sample size, we used the Bootstrap approach to quantitatively control and check the stability of the results by random sampling with replacement.

We reported the R^2, root mean squared error ($RMSE$), and mean absolute error (MAE) for each predictive model in each setting (see Table 6.4). According

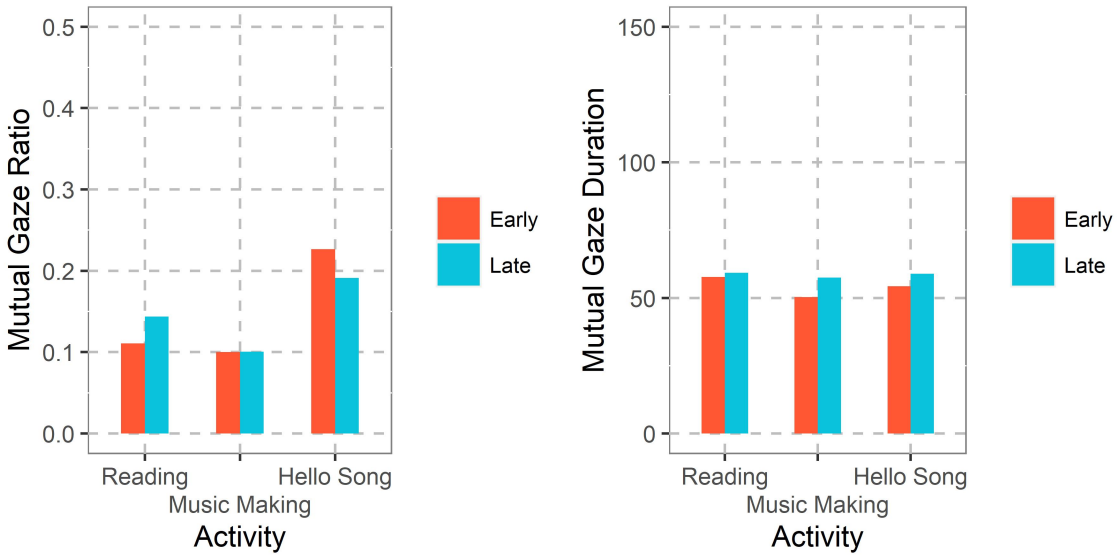


Figure 6.7: The bar chart of mutual gaze ratio and mutual gaze duration in different activities across early and late sessions. No significant social gaze improvement was found. Mutual gaze ratio is in [0–1] range.

to the results, predictions with mutual gaze-related measures got a lower loss than the predictions only based on participant profile data in all models. The random forest regression model with mutual gaze-related measures achieved the best prediction performance ($MAE = 0.348, RMSE = 0.447, R^2 = 0.620$).

6.4 Discussion

While there have not been many studies that quantified social behavior of children with autism using deep learning methods in video-based autism intervention contexts, we here proposed a social visual behavior analytical approach for children in autism using mutual gaze detection. We detected the mutual gaze behaviors by adopting a state-of-the-art gaze detection method [6] and generated mutual gaze-related measures, mutual gaze ratio and mutual gaze duration, based on the detection outcomes. We compared the outcomes of the framework with a manually-coded ground truth measure of social gaze (human-coded ratio) to evaluate the effectiveness of our framework and analyzed the children's social visual behavior using mutual gaze-related

Table 6.4: Summary of social visual behavior score prediction model performance in the ablation experiment.

	Prediction Setting 1			Prediction Setting 2		
Model	MAE	$RMSE$	R^2	MAE	$RMSE$	R^2
RF	**0.348**	**0.447**	**0.620**	0.386	0.464	0.578
Lasso	0.395	0.486	0.558	0.408	0.505	0.494
SVM	0.428	0.513	0.493	0.449	0.530	0.481
GBT	0.455	0.566	0.373	0.482	0.596	0.369
MLP	0.538	0.623	0.241	0.595	0.671	0.159

Prediction Setting 1: prediction based on the mutual gaze ratio and duration, together with the level of functioning and ADOS-2 social affect score.
Prediction Setting 2: prediction only based on the two profile scores, including the level of functioning and ADOS-2 social affect score.
Mutual gaze ratio, mutual gaze duration, level of functioning, and ADOS-2 social affect score are normalized.
Prediction model with setting 1 which includes gaze information has better performance versus prediction model in setting 2 without gaze information (lower values are better for MAE and $RMSE$, and higher value is better for R^2).

measures with an in-house ASD dataset collected from the autism therapy intervention [145].

As we reported in the previous section, our results showed that our framework could successfully report mutual gaze ratio, which was comparable, even replaceable to the human-coded ratio. For different therapy settings, descriptive comparison results of our method matched the results of human-coded measures, which were annotated by subject matter experts. We also found a significant result, which showed a strong positive linear correlation between mutual gaze ratio and human-coded ratio. Compared with typical time-consuming hand-coding approaches to examine the visual performance of children with autism, our method was able to effectively and efficiently extract mutual gaze features during the whole training process with a standard camera.

During the social visual behavior analytics, we examined group setting differences, activity-related within-group effects, and intervention-related changes in social attention. The results showed that there was no significant difference in children's social attention frequency or duration in different group settings. However, in the same group, different activities had significant effects on the frequency of children's social

gaze behaviors. Children in the Play Therapy group performed mutual gaze behavior more frequently during the "Hello Song" activity than the "Music Making" activity. This may be caused by the fact that children with autism focus more on objects used in training, including instruments such as drums and xylophones rather than humans. Given the easy access to objects, children may have engaged in a greater visual fixation on objects. Moreover, the practice of complex drumming and xylophone patterns required sustained monitoring of the instruments. These findings are consistent with the results of the "Reading" activity, in which children with autism also focus on the books when they are reading. Thus, as many researchers reported, object-based activities may reduce children's social visual behaviors during the autism training process [217, 218, 219, 145]. After comparing the mutual gaze ratio and duration between early and late sessions, no significant improvement in social visual performance of children with autism was found after the therapy sessions. This is consistent with the findings of Srinivasan et al. [15] that assessed social gaze for the entire session for early and late sessions based on data coded by human experts. They too reported activity-related differences with great socially-directed gaze during musical play therapy sessions. As is reported in past studies, it is difficult to change levels of sustained social attention in children through short-term behavioral interventions [145, 220, 221].

By examining the prediction models with and without mutual gaze-related measures, we found that the models with mutual gaze ratio and duration have better prediction performance than the models only based on participant profile data. In this case, our mutual gaze-related measures derived from the mutual gaze detection framework can be an important feature to predict the social visual behavior score and be used for social visual behavior evaluation. Among five different regression models, random forest regression model with our mutual gaze-related measures can provide best prediction performance with the lowest loss. Since the social gaze behaviors are scored based on various social gaze patterns, to reduce the loss, more social gaze patterns need to be collected, measured, and used for predictions.

6.5 Conclusion

Capturing children's gaze interaction during the therapeutic training activities, particularly about the mutual gaze towards the trainers, can reveal valuable social visual information; thus, analyzing such behaviors is highly useful for effective therapy treatments and evaluations. In this chapter, we analyzed the social visual behavior performance of the children with autism from the in-house ASD dataset using advanced mutual gaze detection framework [6]. The effectiveness of the framework was validated by comparing the outcome with the social visual behavior measure hand-coded by subject matter experts. According to the social visual behavior analysis, we found that the social attention patterns of children may be affected by certain contexts (with or without learning tools) during the short-term behavioral interventions. Our ablation experiment showed that mutual gaze measures can be a powerful feature for social visual behavior score prediction. The creation and extraction of mutual gaze features can provide valuable information for more accurate analysis and prediction model development. Our findings have implications in social interaction analysis in education technology, therapy evaluation, and intervention design, by offering an analytical approach with a novel assessment tool for social behaviors based on gaze information. Beyond the autism therapy context, our method can be applied to other application scenarios that need reliable automatic social behavior analysis from videos involving human-human or human-virtual agent interactions.

There are still certain limitations in our work as well, which can help us expand on this research in the future. The framework [6] we adopted in this chapter detected the mutual gaze features based on head tracks, which may not be able to capture the correct gaze information for specific situations, e.g., when the eye gaze is not matched with the head direction, e.g., children with ASD could be using peripheral vision and non-foveal vision for tracking individuals; when the head is occluded; when the video record have a low resolution or zoom in/out capturing. To overcome this issue, we plan to investigate and develop more accurate and robust gaze detection methods. We also consider reducing the bias in measurement, annotation, and sampling [222]. For further

implementation, we recommend clinicians and researchers to capture the video in a fix camera position and focus for better gaze detection. This work was still a preliminary study with a limited number of samples. Like most autism studies, this sample was biased to males given the greater ASD prevalence in males. Given the small sample size of this study, we acknowledge that the effect sizes calculated in this study are imprecise with large confidence intervals. There is still a large subset of collected data that we need to include in our future work, which was excluded from the current study due to the absence of social visual behavior annotations. Further evaluation and discussion with clinical experts would be invaluable to design and develop better gaze detection or other social behavior evaluation approaches in practice. In the future work, we aim to improve our framework via semi-supervised and unsupervised machine learning methods to analyze those video recordings without annotation [223, 222]. Moreover, we plan to investigate multi-gaze features and multimodal features, including gaze sharing and following [3, 48], facial expressions [209], and body gestures [210, 211] for framework development and comprehensive interpretation of social visual behaviors.

Chapter 7

CLOSING REMARKS

7.1 Conclusion

In this dissertation, our goal was to contribute to non-verbal behavior analysis research in social and educational contexts. We have been working on automatic non-verbal social features extraction and learning analytics using computer vision methods with deep learning-based frameworks in various educational settings. Firstly, we investigated the physical proximity between collaborating learners using the position information detected from the images by the object detection approach. For object detection, we adopted the deep learning-based Mask R-CNN framework to detect the position of each collaborator with a bounding box and a confidence score to show how likely the detected object is a person. We then generated the measure of Level of Collaboration to represent the overlapping area ratio of the detected bounding boxes of collaborators and used it for collaboration analyses in anatomy learning intervention. According to our knowledge, this was the first work to apply the deep learning-based computer vision method to collaborative learning analytics in co-located learning settings. This work also showed the possibility of using computer vision methods with deep learning-based frameworks to solve new educational problems. However, this method was only suitable for activities-based co-located learning contexts, and camera occlusion was one of the main challenges we faced.

With acknowledgment of our previous work's limitations, we then mainly focused on the social gaze behaviors during the interactive anatomy learning process as the non-verbal cues, which do not require any position change or body movement. JVA, as one of the typical social visual behaviors, was automatically estimated by examining the overlap of the focus points of collaborators predicted by the Gaze Following

framework [3] with our hand-annotated head positions from the images. The Gaze Following framework took the hand-coded coordinates of the head positions and the raw images as the input, predicted the possible focus points of each annotated head based on the detected gaze direction, and outputted the position of each predicted focus point with the heat map. We generated the JVA ratio measure to represent the frequency of joint attention during the learning process and analyzed the collaborative learning performance across different group settings, gender compositions, and team post-test scores in the expanded anatomy learning intervention. This work provided a deep learning-based framework to automatically estimate JVA features based on the image records captured by the standard camera and analyzed collaborative learning performance based on JVA frequency, team post-test score, and other demographic information. This work also showed that gaze signals were suitable non-verbal social behavior measures in learning analytics.

Besides the image-based records, we also investigated non-verbal feature extraction using video records. Mutual attention, known as people looking at each other, is another social visual behavior we are interested in. We used a three-branch head tracking framework [5] for automatic mutual gaze detection based on the in-house ASD dataset collected from an autism physical therapy intervention. The framework took the head detection outcomes as the input, detected mutual gaze behaviors between each pair within the scenes from encoded head position tracks and head maps, and outputted mutual gaze scores for each frame. The mutual gaze score represented the possibility that people were looking at each other in the scene. We generated the mutual gaze ratio based on the mutual gaze scores as a novel measurement. We evaluated our framework using descriptive comparisons, correlation analysis, and regression prediction compared to the ground truth hand-coded social visual behavior score. This work demonstrated the feasibility of applying the mutual gaze detection method to video-based therapy training records. The mutual gaze ratio generated from the detection outcomes was comparable to the social visual behavior score hand-coded by the therapy experts.

Furthermore, we used an advanced head tracking framework [6] to detect the mutual gaze features with higher accuracy. We generated mutual gaze ratio and duration measures from the framework outcomes and evaluated the framework by comparing the mutual gaze ratio with the human-coded ratio. We then analyzed the social visual behaviors of children with autism based on the expanded in-house ASD dataset across different group settings, therapy activities, and therapy sessions. We also used the mutual gaze ratio and duration with the ADOS-2 social affect score [175] and the level of functioning for social visual behavior score prediction. We applied different machine learning regression models, including random forest, support vector regression, Lasso regression, gradient boosting trees regression, and multi-layer perceptron regression (the random forest model had the best performance on the prediction with mutual gaze features). This work provided a new solution for social visual behavior analytics of the autism therapy training process using detected mutual gaze features from the video recordings captured by the standard cameras. This work also showed that mutual gaze behavior could generate reasonable non-verbal social behavior measures in learning analytics, especially in special education contexts.

Overall, our work provided the possibility and reliability of automatically coding, measuring, and evaluating non-verbal interactions in educational environments using deep learning-based frameworks. These works mainly focused on leveraging state-of-the-art deep learning-based frameworks and computer vision methods to extract specific non-verbal cues from visual data as objective measures for further learning analytics. We created different social-behavioral measures inspired by deep learning methods and compared/evaluated these measures with the ground truth hand-annotated data from subject-matter experts. Based on the results, our approaches were successful in providing objective measures for physical proximity, mutual gaze, or JVA, and overall, information on human-human social interactions during the educational process, which were comparable to the hand-coding annotations.

According to the results of our learning analytics, we found both physical proximity and JVA ratio were promising measures in the successful distinction of the study

conditions, and JVA ratio was the more practical and general one for most educational scenarios. During the learning process, collaborators in mobile learning settings or augmented reality settings more frequently got closer and shared attention than those in the standard setting with textbooks. We found a significant positive linear association with a strong relationship between the JVA ratio and the team post-test score in the anatomy learning intervention. This was in agreement with the hypothesis of our study: students who shared gaze with their teammates for a longer time on the learning task were more likely to obtain higher scores in their post-activity knowledge tests. 3D models on the tablet screen or AR system were more attractive for students and helped them easily collaborate and complete the learning tasks with high accuracy. Our study also provided further evidence that 3D visualization technologies increased students' engagement and improved their knowledge retention in human anatomy learning. For gender compositions of the collaborators, no significant difference was observed, neither for the JVA ratio nor for the team post-test score.

For special education contexts, especially for autism therapy training interventions, we examined group setting differences, activity-related within-group effects, and intervention-related changes in social attention. The results showed no significant difference in children's social attention frequency or duration in different group settings. However, in the same group setting, different activities significantly affected the frequency of children's social gaze behaviors. Children with autism perform mutual gaze behavior more frequently during the activities without training tools than the activities with training tools. This might be caused by children participants focusing more on objects used in training, such as instruments, books, etc. Given the easy access to objects, children might engage in a more significant visual fixation on objects. Moreover, complex-task patterns might require sustained monitoring of the objects. Therefore, object-based activities might reduce children participants' social visual behaviors during the training process. After comparing the mutual gaze ratio and duration between early and late sessions, no significant improvement in social visual performance of children participants was found after the therapy sessions. It was challenging to change

levels of sustained social attention in children through short-term behavioral interventions. By examining the prediction models with and without mutual gaze-related measures, we found that the models with mutual gaze ratio and duration had better prediction performance than the models only based on participant profile data. In this case, our mutual gaze-related measures derived from the mutual gaze detection framework could be an essential feature in social visual behavior score prediction and be used for social visual behavior evaluation. Random forest regression with our measures could provide the best prediction performance with the lowest loss among five different regression models.

In conclusion, our methods can automatically recognize non-verbal social behaviors from both image-based and video-based records of the learning process and investigate students' interactive learning performance in normal and special education contexts across different settings. Our approaches and findings have implications in various educational technology, autism therapy analysis, and intervention design fields by offering analytical approaches with novel assessment tools for non-verbal social behaviors based on computer vision methods with deep learning-based frameworks. Our non-verbal social behavior accessing frameworks can be applied to other application scenarios that need reliable automatic social behavior analysis from images or videos involving human-human or human-virtual agent interactions.

There are still certain limitations in our work, which can help us expand on this research in the future. The object detection with the physical proximity measure can only be applied to the co-located collaboration scenarios, which require position changes and body movements. Our frameworks for focus point prediction and mutual gaze detection are based on head information, which may not be able to capture the correct gaze information for specific situations, e.g., in the case that the eye gaze is not matched with the head direction or for people with eye movement disorders. For implementation, we use the pre-trained frameworks for feature extraction due to the lack of annotated ground truth data, which may result in a lower accuracy than the results of models trained with our data. We also need to consider reducing the bias in measurement,

annotation, and sampling when we use deep learning-based frameworks and machine learning models [222]. The focus of our learning analytics is to understand learning performance in co-located situations, and gaze features were identified during the post-analysis process. Our analytics are only based on students or child-trainer pairs in dyadic interactions. This work was still a preliminary study with a limited number of samples. Like most autism studies, recruitment of children with autism was not easy, which made the balancing of samples across different study conditions very challenging. Given the small sample size of this study, we acknowledge that the effect sizes calculated in this dissertation are imprecise with large confidence intervals. Further evaluation and discussion with subject-matter experts (autism therapists) would be invaluable to designing and developing better gaze detection or other social behavior evaluation approaches in practice.

7.2 Future Work

Our work has shown that gaze cues are an observable, reliable, and efficient measure used for non-verbal interaction assessment and evaluation in the educational environment. However, the single gaze feature cannot represent well all the social gaze patterns of the interactions. Different gaze behaviors such as gaze alignment, mutual gaze attentiveness, and other gaze patterns may occur in the same learning process. We anticipate that, compared with single gaze behavior analyses, multi-gaze features analyses can provide more valuable and comprehensive information on the social gaze patterns in human-human interactions. While research on using multi-gaze features in the uni-modal using deep learning for learning analytics is still limited, we plan to combine JVA and mutual gaze features for further social visual gaze behavior pattern analyses in our future work. In the planned research, the output of the head detection step will be used as the input of the gaze point prediction framework instead of hand annotation in Chapter 4 to save more time and effort for experts and educators.

Researchers have shown that multimodal learning analytics could offer new insights into students' learning trajectories [224, 223, 130]. Besides the gaze signals,

we plan to collect multimodal non-verbal behavioral features based on deep learning frameworks, such as gaze sharing and following [39, 47], facial expressions [209, 153], and body gestures [210, 211] for a more comprehensive interpretation of social visual behaviors, including social engagement and social anxiety [212] of children in the spectrum. Multimodal data collection and analysis techniques can bring novel methods to understand students' learning performance during the group discussion, interaction with peers, and actions in both the digital and the physical worlds [224]. Additionally, we would like to compare the accuracy of machine learning-based models of non-verbal social behavior prediction induced with several different combinations of input modalities. These ablation experiments will be conducted to evaluate predictive models that use the video data collected from the standard camera at the fixed location that are less disruptive to learning.

Non-verbal feature extraction in our current work is based on the pre-trained models due to the limited annotated ground truth. In the future work, we aim to extend our framework via semi-supervised and unsupervised machine learning methods to analyze the non-annotated video recordings. We also aim to hand-annotate part of our collected data based on the object and social feature categories we need, including learning tools and students' non-verbal behaviors (head poses/facial expressions/body gestures) as the ground truth, and re-train our model on the training dataset for higher accuracy. To improve the feature extraction performance, we would like to investigate the latest state-of-the-art frameworks in their own area.

Different sample sizes and the balance of the data composition may draw different results and conclusions. We plan to extend the sample size for further learning performance analytics with more students' profile information as another dimension to perform collaboration evaluations such as age, gender, and level of social and communication skills.

Non-verbal social behavior analytics in collaborative learning and autism therapy intervention is just part of our purposes. We believe that most of the work presented in this dissertation can be directly extended to other scenarios. In future work,

it will be ideal to evaluate the effectiveness of adopting our method in other social learning scenarios. In further planned research, a data set with team sizes larger than two would better illustrate our idea and validate our findings and would be suitable for more educational scenarios.

BIBLIOGRAPHY

[1] Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014.

[2] Massimiliano Patacchiola and Angelo Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 2017.

[3] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: Revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.

[6] Manuel Jesus Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net++: revisiting people looking at each other in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[7] Deepika Phutela. The importance of non-verbal communication. *IUP Journal of Soft Skills*, 9(4):43, 2015.

[8] Zhang Guo and Roghayeh Barmaki. Collaboration analysis using object detection. In *Proceedings of the 2019 Educational Data Mining*, 2019.

[9] Ali Derakhshan. The predictability of turkman students' academic engagement through persian language teachers' nonverbal immediacy and credibility. *Journal of Teaching Persian to Speakers of Other Languages*, 10(21):3–26, 2021.

[10] Tariq Hussain, Asmaa Azeem, and Nisar Abid. Examining the correlation between university students' perceived teacher immediacy and their motivation. *Psychology and Education Journal*, 58(1):5809–5820, 2021.

[11] Sukris Sutiyatno. The effect of teacher's verbal communication and non-verbal communication on students' english achievement. *Journal of Language Teaching and Research*, 9(2):430–437, 2018.

[12] Gill Westland. *Verbal and non-verbal communication in psychotherapy*. WW Norton and Company, 2015.

[13] Lauren G Collins, Anne Schrimmer, James Diamond, and Janice Burke. Evaluating verbal and non-verbal communication skills, in an ethnogeriatric osce. *Patient education and counseling*, 83(2):158–162, 2011.

[14] Han-Yu Sung and Gwo-Jen Hwang. A collaborative game-based learning approach to improving students' learning performance in science courses. *Computers and Education*, 63:43–51, 2013.

[15] Sudha M Srinivasan, Inge-Marie Eigsti, Timothy Gifford, and Anjana N Bhat. The effects of embodied rhythm and robotic interventions on the spontaneous and responsive verbal communication skills of children with autism spectrum disorder (asd): A further outcome of a pilot randomized controlled trial. *Research in Autism Spectrum Disorders*, 27:73–87, 2016.

[16] Roghayeh Barmaki, Kevin Yu, Rebecca Pearlman, Richard Shingles, Felix Bork, Greg M Osgood, and Nassir Navab. Enhancement of anatomical education using augmented reality: An empirical study of body painting. *Anatomical sciences education*, 2019.

[17] Karina Huang, Tonya Bryant, and Bertrand Schneider. Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. *International Educational Data Mining Society*, 2019.

[18] Fleur Ruth Prinsen, Monique LL Volman, and Jan Terwel. Gender-related differences in computer-mediated communication and computer-supported collaborative learning. *Journal of Computer Assisted Learning*, 23(5):393–409, 2007.

[19] Bertrand Schneider, Kshitij Sharma, Sebastien Cuendet, Guillaume Zufferey, Pierre Dillenbourg, and Roy Pea. Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning*, 13(3):241–261, 2018.

[20] Sara De Freitas and Mark Griffiths. Online gaming as an educational tool in learning and training. *British Journal of Educational Technology*, 38(3):535–537, 2007.

[21] Qing Li, Rynson WH Lau, Timothy K Shih, and Frederick WB Li. Technology supports for distributed and collaborative learning over the internet. *ACM Transactions on Internet Technology (TOIT)*, 8(2):1–24, 2008.

[22] Frederico Menine Schaf, Dieter Müller, F Wilhelm Bruns, Carlos Eduardo Pereira, and H-H Erbe. Collaborative learning and engineering workspaces. *Annual Reviews in Control*, 33(2):246–252, 2009.

[23] Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 126–146. IEEE, 2019.

[24] Michael Argyle, Roger Ingham, Florisse Alkema, and Margaret McCallin. The different functions of gaze. *Semiotica*, 7(1):19–32, 1973.

[25] Bertrand Schneider and Roy Pea. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. In *Educational media and technology yearbook*, pages 99–125. Springer, 2017.

[26] Vincent Van Rheden, Bernhard Maurer, Dorothé Smit, Martin Murer, and Manfred Tscheligi. Laserviz: Shared gaze in the co-located physical world. In *Proceedings of the Eleventh International Conference on Tangible, Embedded, and Embodied Interaction*, pages 191–196, 2017.

[27] Sami Pietinen, Roman Bednarik, and Markku Tukiainen. Shared visual attention in collaborative programming: a descriptive analysis. In *proceedings of the 2010 ICSE workshop on cooperative and human aspects of software engineering*, pages 21–24, 2010.

[28] Hidde van der Meulen, Petra Varsanyi, Lauren Westendorf, Andrew L Kun, and Orit Shaer. Towards understanding collaboration around interactive surfaces: Exploring joint visual attention. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 219–220, 2016.

[29] Basil Wahn, Jessika Schwandt, Matti Krüger, Daina Crafa, Vanessa Nunnendorf, and Peter König. Multisensory teamwork: using a tactile or an auditory display to exchange gaze information improves performance in joint visual search. *Ergonomics*, 59(6):781–795, 2016.

[30] Benjamin Zablotsky, Lindsey I Black, Matthew J Maenner, Laura A Schieve, Melissa L Danielson, Rebecca H Bitsko, Stephen J Blumberg, Michael D Kogan, and Coleen A Boyle. Prevalence and trends of developmental disabilities among children in the united states: 2009–2017. *Pediatrics*, 144(4), 2019.

[31] S Mohammad Mavadati, Huanghao Feng, Anibal Gutierrez, and Mohammad H Mahoor. Comparing the gaze responses of children with autism and typically developed individuals in human-robot interaction. In *Proceedings of 2014 IEEE-RAS International Conference on Humanoid Robots*, pages 1128–1133. IEEE, 2014.

[32] Geraldine Dawson, Karen Toth, Robert Abbott, Julie Osterling, Jeff Munson, Annette Estes, and Jane Liaw. Early social attention impairments in autism: social orienting, joint attention, and attention to distress. *Developmental Psychology*, 40(2):271, 2004.

[33] Sally Ozonoff, Ana-Maria Iosif, Fam Baguio, Ian C Cook, Monique Moore Hill, Ted Hutman, Sally J Rogers, Agata Rozga, Sarabjit Sangha, Marian Sigman, et al. A prospective study of the emergence of early behavioral signs of autism. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(3):256–266, 2010.

[34] Julie A Osterling, Geraldine Dawson, and Jeffrey A Munson. Early recognition of 1-year-old infants with autism spectrum disorder versus mental retardation. *Development and Psychopathology*, 14(2):239–251, 2002.

[35] Paul Yoder, Wendy L Stone, Tedra Walden, and Elizabeth Malesa. Predicting social impairment and asd diagnosis in younger siblings of children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 39(10):1381–1391, 2009.

[36] Lindsey Sterling, Geraldine Dawson, Annette Estes, and Jessica Greenson. Characteristics associated with presence of depressive symptoms in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 38(6):1011–1018, 2008.

[37] Jon Baio, Lisa Wiggins, Deborah L Christensen, Matthew J Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, Walter Zahorodny, Cordelia Robinson Rosenberg, Tiffany White, et al. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2014. *MMWR Surveillance Summaries*, 67(6):1, 2018.

[38] Cristina Palmero, Elsbeth A van Dam, Sergio Escalera, Mike Kelia, Guido F Lichtert, Lucas PJJ Noldus, Andrew J Spink, and Astrid van Wieringen. Automatic mutual gaze detection in face-to-face dyadic interaction videos. In *Proceedings of Measuring Behavior*, volume 1, page 2, 2018.

[39] Zhang Guo and Roghayeh Barmaki. Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction. *Australasian Journal of Educational Technology*, 36(6):53–71, 2020.

[40] Chidchanok Thepsoonthorn, Takahiro Yokozuka, Jinhwan Kwon, Robin Miao Sin Yap, Shunsuke Miura, Ken-ichiro Ogawa, and Yoshihiro Miyake. Look at you, look at me: Detection and analysis of mutual gaze convergence in face-to-face interaction. In *Proceedings of 2015 IEEE/SICE International Symposium on System Integration*, pages 581–586. IEEE, 2015.

[41] Daniel C Richardson and Rick Dale. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29(6):1045–1060, 2005.

[42] David Wendler. One-time general consent for research on biological samples: is it compatible with the health insurance portability and accountability act? *Archives of internal medicine*, 166(14):1449–1452, 2006.

[43] Anjana N Bhat, Rebecca J Landa, and James C Galloway. Current perspectives on motor functioning in infants, children, and adults with autism spectrum disorders. *Physical Therapy*, 91(7):1116–1129, 2011.

[44] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D Abowd, and James M Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 699–704, 2012.

[45] Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M Rehg. Detecting bids for eye contact using a wearable camera. In *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.

[46] Sidrah Liaqat, Chongruo Wu, Prashanth Reddy Duggirala, Sen-ching Samson Cheung, Chen-Nee Chuah, Sally Ozonoff, and Gregory Young. Predicting asd diagnosis in children with synthetic and image-based eye gaze data. *Signal Processing: Image Communication*, 94:116198, 2021.

[47] Jicheng Li and Roghayeh Barmaki. Trends in virtual and augmented reality research: A review of latest eye tracking research papers and beyond. *Preprints*, 2019.

[48] Zhang Guo, Kangsoo Kim, Anjana Bhat, and Roghayeh Barmaki. An automated mutual gaze detection framework for social behavior assessment in therapy for children with autism. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 444–452, 2021.

[49] Mark E Comadena, Stephen K Hunt, and Cheri J Simonds. The effects of teacher clarity, nonverbal immediacy, and caring on student motivation, affective and cognitive learning. *Communication Research Reports*, 24(3):241–248, 2007.

[50] Jenny Mackay. *Coat of many pockets: Managing classroom interactions*. Aust Council for Ed Research, 2006.

[51] Daniela Sime. What do learners make of teachers' gestures in the language classroom? 2006.

[52] Jun Liu. *Asian students' classroom communication patterns in US universities: An emic perspective*. Greenwood Publishing Group, 2001.

[53] Marian L Houser and Ann Bainbridge Frymier. The role of student characteristics and teacher behaviors in students' learner empowerment. *Communication Education*, 58(1):35–53, 2009.

[54] Melinda Lincoln. *Conflict resolution communication: Patterns promoting peaceful schools*. Scarecrow Press, 2002.

[55] Sean Neill. *Classroom nonverbal communication*. Routledge, 2017.

[56] Sylvia Helmer, Catherine Eddy, and Catherine L Eddy. *Look at me when I talk to you: ESL learners in non-ESL classrooms*, volume 39. Pippin Publishing Corporation, 2003.

[57] Johann Le Roux. Effective educators are culturally competent communicators. *Intercultural Education*, 13(1):37–48, 2002.

[58] John J Okon. Role of non-verbal communication in education. *Mediterranean Journal of Social Sciences*, 2(5):35–35, 2011.

[59] Erik P Bucy and Patrick Stewart. The personalization of campaigns: Nonverbal cues in presidential debates. In *Oxford Research Encyclopedia of Politics*. 2018.

[60] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[61] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[62] Erik P Bucy. The look of losing, then and now: Nixon, obama, and nonverbal indicators of opportunity lost. *American Behavioral Scientist*, 60(14):1772–1798, 2016.

[63] Anjana N Bhat. Motor impairment increases in children with autism spectrum disorder as a function of social communication, cognitive and functional impairment, repetitive behavior severity, and comorbid diagnoses: A spark study report. *Autism Research*, 14(1):202–219, 2021.

[64] Anjana Narayan Bhat. Is motor impairment in autism spectrum disorder distinct from developmental coordination disorder? a report from the spark study. *Physical Therapy*, 100(4):633–644, 2020.

[65] P Alex Dow, Lada A Adamic, and Adrien Friggeri. The anatomy of large facebook cascades. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[66] Joel B Hagen. The origins of bioinformatics. *Nature Reviews Genetics*, 1(3):231–236, 2000.

[67] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Comput. Linguist*, 35(2):311–312, 2009.

[68] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[69] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[70] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[71] Kshitij Bantupalli and Ying Xie. American sign language recognition using deep learning and computer vision. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4896–4899. IEEE, 2018.

[72] Isha Talegaonkar, Kalyani Joshi, Shreya Valunj, Rucha Kohok, and Anagha Kulkarni. Real time facial expression recognition using deep learning. In *Proceedings of International Conference on Communication and Information Processing (ICCIP)*, 2019.

[73] Pierre Dillenbourg. What do you mean by collaborative learning?, 1999.

[74] Teresa Monahan, Gavin McArdle, and Michela Bertolotto. Virtual reality for collaborative e-learning. *Computers and Education*, 50(4):1339–1353, 2008.

[75] Christian Dalsgaard. Social software: E-learning beyond learning management systems. *European Journal of Open, Distance and e-learning*, 9(2), 2006.

[76] Rateeba Algoufi. Using tablet on education. *World Journal of Education*, 6(3):113–119, 2016.

[77] Wing Sum Cheung and Khe Foon Hew. A review of research methodologies used in studies on mobile handheld devices in k-12 and higher education settings. *Australasian Journal of Educational Technology*, 25(2):153–183, 2009.

[78] Marwan F Abu-Hijleh. The place of anatomy in medical education: Guide supplement 41.1–viewpoint. *Medical teacher*, 32(7):601–603, 2010.

[79] Paul G McMenamin. Body painting as a tool in clinical anatomy teaching. *Anatomical sciences education*, 1(4):139–144, 2008.

[80] John Dewey. Democracy in education. *The elementary school teacher*, 4(4):193–204, 1903.

[81] Alan Liu, Frank Tendick, Kevin Cleary, and Christoph Kaufmann. A survey of surgical simulation: applications, technology, and education. *Presence: Teleoperators and virtual environments*, 12(6):599–614, 2003.

[82] César A Collazos, Luis A Guerrero, José A Pino, and Sergio F Ochoa. Evaluating collaborative learning processes. In *International Conference on Collaboration and Technology*, pages 203–221. Springer, 2002.

[83] Amy Soller, Alan Lesgold, Frank Linton, and Brad Goodman. What makes peer interaction effective? modeling effective communication in an intelligent cscl. In *Proceedings of the 1999 AAAI fall symposium: Psychological models of communication in collaborative systems*, pages 116–123. Cape Cod, 1999.

[84] Amy Soller and Alan Lesgold. Modeling the process of collaborative learning. In *The role of technology in CSCL*, pages 63–86. Springer, 2007.

[85] César A Collazos, Luis A Guerrero, José A Pino, Stefano Renzi, Jane Klobas, Manuel Ortega, Miguel A Redondo, and Crescencio Bravo. Evaluating collaborative learning processes using system-based measurement. *Educational Technology and Society*, 10(3):257–274, 2007.

[86] Amruta Chavan, Dipali Bendale, Radha Shimpi, and Pradnya Vikhar. Object detection and recognition in images. *International Journal of Computing and Technology*, 3(3):148–151, 2016.

[87] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[88] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[89] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[90] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[91] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[92] Kaissar Yammine and Claudio Violato. A meta-analysis of the educational effectiveness of three-dimensional visualization technologies in teaching anatomy. *Anatomical sciences education*, 8(6):525–538, 2015.

[93] Nikola Golenhofen, Felix Heindl, Claudia Grab-Kroll, David AC Messerer, Tobias M Böckers, and Anja Böckers. The use of a mobile learning tool by medical students in undergraduate anatomy and its effects on assessment outcomes. *Anatomical sciences education*, 13(1):8–18, 2020.

[94] Robson R Lemos, Cristiane Meneghelli Rudolph, Arthur V Batista, Karolini R Conceição, Poliana F Pereira, Bruna S Bueno, Patricia J Fiuza, and Samira S Mansur. Design of a web3d serious game for human anatomy education: A web3d game for human anatomy education. In *Handbook of research on immersive digital games in educational environments*, pages 586–611. IGI Global, 2019.

[95] HS Maresky, A Oikonomou, I Ali, N Ditkofsky, M Pakkal, and B Ballyk. Virtual reality and cardiac anatomy: Exploring immersive three-dimensional cardiac imaging, a pilot study in undergraduate medical anatomy education. *Clinical Anatomy*, 32(2):238–243, 2019.

[96] Jennifer NA Silva, Michael Southworth, Constantine Raptis, and Jonathan Silva. Emerging applications of virtual reality in cardiovascular medicine. *JACC: Basic to Translational Science*, 3(3):420–430, 2018.

[97] Matthew Hackett and Michael Proctor. The effect of autostereoscopic holograms on anatomical knowledge: a randomised trial. *Medical education*, 52(11):1147–1155, 2018.

[98] Jan-Maarten Luursema, Willem B Verwey, Piet AM Kommers, Robert H Geelkerken, and Hans J Vos. Optimizing conditions for computer-assisted anatomical learning. *Interacting with Computers*, 18(5):1123–1138, 2006.

[99] Jan-Maarten Luursema, Willem B Verwey, Piet AM Kommers, and Jan-Henk Annema. The role of stereopsis in virtual anatomical learning. *Interacting with Computers*, 20(4-5):455–460, 2008.

[100] Luis Fernandez-Sanz and Sanjay Misra. Analysis of cultural and gender influences on teamwork performance for software requirements analysis in multinational environments. *IET software*, 6(3):167–175, 2012.

[101] Jürgen Wegge, Carla Roth, Barbara Neubach, Klaus-Helmut Schmidt, and Ruth Kanfer. Age and gender diversity as determinants of performance and health in a public organization: the role of task complexity and group size. *Journal of Applied Psychology*, 93(6):1301, 2008.

[102] Patrick Rabbitt, Christopher Donlan, Peter Watson, Lynn McInnes, and Nuala Bent. Unique and interactive effects of depression, age, socioeconomic advantage, and gender on cognitive performance of normal healthy older people. *Psychology and aging*, 10(3):307, 1995.

[103] K Warner Schaie and Sherry L Willis. Age difference patterns of psychometric intelligence in adulthood: generalizability within and across ability domains. *Psychology and aging*, 8(1):44, 1993.

[104] Julia B Bear and Anita Williams Woolley. The role of gender in team collaboration and performance. *Interdisciplinary science reviews*, 36(2):146–153, 2011.

[105] Maria De Paola, Francesca Gioia, and Vincenzo Scoppa. Teamwork, leadership and gender. Technical report, IZA Discussion Papers, 2018.

[106] Alice H Eagly and Linda L Carli. The female leadership advantage: An evaluation of the evidence. *The leadership quarterly*, 14(6):807–834, 2003.

[107] Lorelle A Meadows, Denise Sekaquaptewa, Marie C Paretti, Alice L Pawley, Shawn S Jordan, Debbie Chachra, and Adrienne Minerick. Interactive panel: Improving the experiences of marginalized students on engineering design teams. In *2015 ASEE Annual Conference and Exposition*, pages 26–1007, 2015.

[108] Peter Meiksins, P Layne, E Camargo, and K Snead. Women in engineering: A review of the 2014 literature. *PREPARE TO PRACTICE CURIOSITY*, 4, 2013.

[109] Jan Andersson. Net effect of memory collaboration: How is collaboration affected by factors such as friendship, gender and age? *Scandinavian journal of Psychology*, 42(4):367–375, 2001.

[110] Susan C Herring. *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, volume 39. John Benjamins Publishing, 1996.

[111] Lasse Lipponen. The challenges for computer supported collaborative learning in elementary and secondary level: Finnish perspectives. 1999.

[112] Sandra Okita, Jeremy Bailenson, and Daniel Schwartz. Mere belief of social action improves complex learning. 2008.

[113] Noreen M Webb, Jonathan D Troper, and Randy Fall. Constructive activity and learning in collaborative small groups. *Journal of educational psychology*, 87(3):406, 1995.

[114] Ryan SJD Baker, Kalina Yacef, et al. The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, 1(1):3–17, 2009.

[115] Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning*, 3(3):237–271, 2008.

[116] Sotiris B Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4):331–344, 2012.

[117] Shree Krishna Subburaj, Angela EB Stewart, Arjun Ramesh Rao, and Sidney K D'Mello. Multimodal, multiparty modeling of collaborative problem solving performance. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 423–432, 2020.

[118] Hana Vrzakova, Mary Jean Amon, Angela EB Stewart, and Sidney K D'Mello. Dynamics of visual attention in multiparty collaborative problem solving using multidimensional recurrence quantification analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[119] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 191–199, 2018.

[120] Go Miura and Shogo Okada. Task-independent multimodal prediction of group performance based on product dimensions. In *2019 International Conference on Multimodal Interaction*, pages 264–273, 2019.

[121] Gabriel Murray and Catharine Oertel. Predicting group performance in task-based interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 14–20, 2018.

[122] Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*, pages 31–41, 2016.

[123] Xavier Ochoa, Katherine Chiluiza, Gonzalo Méndez, Gonzalo Luzardo, Bruno Guamán, and James Castells. Expertise estimation based on simple multimodal features. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 583–590, 2013.

[124] Daniel Spikol, Emanuele Ruffaldi, and Mutlu Cukurova. Using multimodal learning analytics to identify aspects of collaboration in project-based learning. Philadelphia, PA: International Society of the Learning Sciences., 2017.

[125] Susan E Brennan, Xin Chen, Christopher A Dickinson, Mark B Neider, and Gregory J Zelinsky. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477, 2008.

[126] Gary Bente, Felix Eschenburg, and Nicole C Krämer. Virtual gaze. a pilot study on the effects of computer simulated gaze in avatar-based conversations. In *International Conference on Virtual Reality*, pages 185–194. Springer, 2007.

[127] Jessica Markus, Peter Mundy, Michael Morales, Christine EF Delgado, and Marygrace Yale. Individual differences in infant skills as predictors of child-caregiver joint attention and language. *Social Development*, 9(3):302–315, 2000.

[128] Christina Whalen and Laura Schreibman. Joint attention training for children with autism using behavior modification procedures. *Journal of Child psychology and psychiatry*, 44(3):456–468, 2003.

[129] Tonya Bryant, Iulian Radu, and Bertrand Schneider. A qualitative analysis of joint visual attention and collaboration with high-and low-achieving groups in computer-mediated learning. 2019.

[130] Yanghee Kim, Cynthia D'Angelo, Francesco Cafaro, Xavier Ochoa, Danielle Espino, Aaron Kline, Eric Hamilton, Seung Lee, Sachit Butail, Lichuan Liu, et al. Multimodal data analytics for assessing collaborative interactions. 2020.

[131] Zeynep Yücel, Albert Ali Salah, Çetin Meriçli, Tekin Meriçli, Roberto Valenti, and Theo Gevers. Joint attention by gaze interpolation and saliency. *IEEE Transactions on cybernetics*, 43(3):829–842, 2013.

[132] Daniel Harari, Joshua B Tenenbaum, and Shimon Ullman. Discovery and usage of joint attention in images. *arXiv preprint arXiv:1804.04604*, 2018.

[133] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.

[134] Adria Recasens Continente, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? 2015.

[135] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015.

[136] American Psychiatric Association et al. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Pub, 2013.

[137] Peter C Mundy. *Autism and Joint Attention: Development, Neuroscience, and Clinical Fundamentals*. Guilford Publications, 2016.

[138] Mayada Elsabbagh, Janice Fernandes, Sara Jane Webb, Geraldine Dawson, Tony Charman, Mark H Johnson, British Autism Study of Infant Siblings Team, et al. Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood. *Biological Psychiatry*, 74(3):189–194, 2013.

[139] Brandon Keehn, Alan J Lincoln, Ralph-Axel Müller, and Jeanne Townsend. Attentional networks in children and adolescents with autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 51(11):1251–1259, 2010.

[140] Heidi Hillman. Child-centered play therapy as an intervention for children with autism: A literature review. *International Journal of Play Therapy*, 27(4):198, 2018.

[141] Jenifer Ware Balch and Dee C Ray. Emotional assets of children with autism spectrum disorder: A single-case therapeutic outcome experiment. *Journal of Counseling and Development*, 93(4):429–439, 2015.

[142] Loretta Gallo-Lopez and Lawrence C Rubin. *Play-Based Interventions for Children and Adolescents with Autism Spectrum Disorders*. Routledge, 2012.

[143] Dee C Ray and Sue C Bratton. What the research shows about play therapy: Twenty-first century update. *Child-centered Play Therapy Research: The Evidence Base for Effective Practice*, pages 3–33, 2010.

[144] Monika Geretsegger, Cochavit Elefant, Karin A Mössler, and Christian Gold. Music therapy for people with autism spectrum disorder. *Cochrane Database of Systematic Reviews*, (6), 2014.

[145] Sudha M Srinivasan, Inge-Marie Eigsti, Linda Neelly, and Anjana N Bhat. The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with autism spectrum disorder (asd): A pilot randomized controlled trial. *Research in Autism Spectrum Disorders*, 27:54–72, 2016.

[146] Kristen L Hess, Michael J Morrier, L Juane Heflin, and Michelle L Ivey. Autism treatment survey: Services received by children with autism spectrum disorders in public school classrooms. *Journal of Autism and Developmental Disorders*, 38(5):961–971, 2008.

[147] Hayoung A Lim and Ellary Draper. The effects of music therapy incorporated with applied behavior analysis verbal behavior approach for children with autism spectrum disorders. *Journal of Music Therapy*, 48(4):532–550, 2011.

[148] Anna Bonnel, Laurent Mottron, Isabelle Peretz, Manon Trudel, Erick Gallun, and Anne-Marie Bonnel. Enhanced pitch sensitivity in individuals with autism: a signal detection analysis. *Journal of Cognitive Neuroscience*, 15(2):226–235, 2003.

[149] Christina Yvonne Jones. *The effects of music therapy frequency on children with autism spectrum disorder (ASD); The therapists point of view*. PhD thesis, Northcentral University, 2021.

[150] Anjana Narayan Bhat and Sudha Srinivasan. A review of "music and movement" therapies for children with autism: embodied interventions for multisystem development. *Frontiers in Integrative Neuroscience*, 7:22, 2013.

[151] Basilio Noris, Jacqueline Nadel, Mandy Barker, Nouchine Hadjikhani, and Aude Billard. Investigating gaze of children with asd in naturalistic settings. *PloS One*, 7(9):e44144, 2012.

[152] Elgiz Bal, Emily Harden, Damon Lamb, Amy Vaughan Van Hecke, John W Denver, and Stephen W Porges. Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of Autism and Developmental Disorders*, 40(3):358–370, 2010.

[153] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. A two-stage multi-modal affect analysis framework for children with autism spectrum disorder. *arXiv preprint arXiv:2106.09199*, 2021.

[154] Patricia L Mirenda, Anne M Donnellan, and David E Yoder. Gaze behavior: A new look at an old problem. *Journal of autism and developmental disorders*, 13(4):397–409, 1983.

[155] Michael Rutter. Diagnosis and definition. In *Autism*, pages 1–25. Springer, 1978.

[156] Lorna Wing and Judith Gould. Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of autism and developmental disorders*, 9(1):11–29, 1979.

[157] Maria Eleonora Minissi, Irene Alice Chicchi Giglioli, Fabrizia Mantovani, and Mariano Alcaniz Raya. Assessment of the autism spectrum disorder based on machine learning and social visual attention: A systematic review. *Journal of Autism and Developmental Disorders*, pages 1–16, 2021.

[158] Terje Falck-Ytter and Claes von Hofsten. How special is social looking in asd: a review. *Progress in brain research*, 189:209–222, 2011.

[159] Meia Chita-Tegmark. Social attention in asd: A review and meta-analysis of eye-tracking studies. *Research in developmental disabilities*, 48:79–93, 2016.

[160] Thomas W Frazier, Mark Strauss, Eric W Klingemier, Emily E Zetzer, Antonio Y Hardan, Charis Eng, and Eric A Youngstrom. A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(7):546–555, 2017.

[161] Roser Cañigueral and Antonia F de C Hamilton. The role of eye gaze during natural social interactions in typical and autistic people. *Frontiers in Psychology*, 10:560, 2019.

[162] Coralie Chevallier, Julia Parish-Morris, Alana McVey, Keiran M Rump, Noah J Sasson, John D Herrington, and Robert T Schultz. Measuring social attention and motivation in autism spectrum disorder using eye-tracking: Stimulus type matters. *Autism Research*, 8(5):620–628, 2015.

[163] Elisabeth AH von dem Hagen and Naomi Bright. High autistic trait individuals do not modulate gaze behaviour in response to social presence but look away more when actively engaged in an interaction. *Autism Research*, 10(2):359–368, 2017.

[164] Fred R Volkmar and Linda C Mayes. Gaze behavior in autism. *Development and Psychopathology*, 2(1):61–69, 1990.

[165] Bradley M Drysdale, Dennis W Moore, Brett E Furlonger, and Angelika Anderson. Gaze patterns of individuals with asd during active task engagement: a systematic literature review. *Review Journal of Autism and Developmental Disorders*, 5(1):1–14, 2018.

[166] Zhong Zhao, Haiming Tang, Xiaobin Zhang, Zhipeng Zhu, Jiayi Xing, Wenzhou Li, Da Tao, Xingda Qu, and Jianping Lu. Characteristics of visual fixation in chinese children with autism during face-to-face conversations. *Journal of Autism and Developmental Disorders*, pages 1–13, 2021.

[167] Peter Mundy, Marian Sigman, and Connie Kasari. Joint attention, developmental level, and symptom presentation in autism. *Development and Psychopathology*, 6(3):389–401, 1994.

[168] Zillah Boraston and Sarah-Jayne Blakemore. The application of eye-tracking technology in the study of autism. *The Journal of physiology*, 581(3):893–898, 2007.

[169] Camila Alviar, Rick Dale, Akeiylah Dewitt, and Christopher Kello. Multimodal coordination of sound and movement in music and speech. *Discourse Processes*, 57(8):682–702, 2020.

[170] Rima-Maria Rahal and Susann Fiedler. Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, 85:103842, 2019.

[171] Gregory Funke, Eric Greenlee, Martha Carter, Allen Dukes, Rebecca Brown, and Lauren Menke. Which eye tracker is right for your research? performance evaluation of several cost variant eye trackers. In *Proceedings of the Human Factors and Ergonomics Society annual meeting*, volume 60, pages 1240–1244. SAGE Publications Sage CA: Los Angeles, CA, 2016.

[172] Benedikt Hosp, Shahram Eivazi, Maximilian Maurer, Wolfgang Fuhl, David Geisler, and Enkelejda Kasneci. Remoteeye: An open-source high-speed remote eye tracker. *Behavior research methods*, 52(3):1387–1401, 2020.

[173] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael 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):205–223, 2000.

[174] Catherine Lord, Michael Rutter, Susan Goode, Jacquelyn Heemsbergen, Heather Jordan, Lynn Mawhood, and Eric Schopler. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders*, 19(2):185–212, 1989.

[175] Katherine Gotham, Andrew Pickles, and Catherine Lord. Standardizing ados scores for a measure of severity in autism spectrum disorders. *Journal of autism and developmental disorders*, 39(5):693–705, 2009.

[176] Sudha M Srinivasan, Maninderjit Kaur, Isabel K Park, Timothy D Gifford, Kerry L Marsh, and Anjana N Bhat. The effects of rhythm and robotic interventions on the imitation/praxis, interpersonal synchrony, and motor performance of children with autism spectrum disorder (asd): A pilot randomized controlled trial. *Autism Research and Treatment*, 2015, 2015.

[177] Brijesh Kumar Baradwaj and Saurabh Pal. Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*, 2012.

[178] Farshid Marbouti, Heidi A Diefes-Dux, and Krishna Madhavan. Models for early prediction of at-risk students in a course using standards-based grading. *Computers and Education*, 103:1–15, 2016.

[179] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[180] Maximilian Riesenhuber and Tomaso Poggio. Models of object recognition. *Nature neuroscience*, 3(11s):1199, 2000.

[181] Elaine Nicpon Marieb and Pamela B Jackson. *Essentials of Human Anatomy and Physiology Laboratory Manual*. Pearson/Benjamin Cummings, 2006.

[182] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[183] Jeremy Ng, Xiao Hu, Miyu Luo, and Sam KW Chu. Relations among participation, fairness and performance in collaborative learning with wiki-based analytics. *Proceedings of the Association for Information Science and Technology*, 56(1):463–467, 2019.

[184] Judith E Innes and David E Booher. Collaborative rationality as a strategy for working with wicked problems. *Landscape and urban planning*, 154:8–10, 2016.

[185] Dimitris Bertsimas and Shubham Gupta. Fairness and collaboration in network air traffic flow management: an optimization approach. *Transportation Science*, 50(1):57–76, 2016.

[186] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.

[187] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 2594–2605, 2017.

[188] Elaine N Marieb. *Essentials of Human Anatomy and Physiology Laboratory Manual*. Pearson Higher Ed, 2015.

[189] Edmar Rezende, Guilherme Ruppert, Tiago Carvalho, Fabio Ramos, and Paulo De Geus. Malicious software classification using transfer learning of resnet-50 deep neural network. In *2017 16th IEEE International Conference on Machine Learning and Applications*, pages 1011–1014. IEEE, 2017.

[190] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9259–9266, 2019.

[191] Yvonne Bruinsma, Robert L Koegel, and Lynn Kern Koegel. Joint attention and children with autism: A review of the literature. *Mental retardation and developmental disabilities research reviews*, 10(3):169–175, 2004.

[192] Barbara Barzansky, Harry S Jonas, and Sylvia I Etzel. Educational programs in us medical schools, 1997-1998. *JAMA*, 280(9):803–808, 1998.

[193] M. Brooks. More women than men enrolled in us med schools for first time. *Medscape*, 2017.

[194] Kelly H Zou, Kemal Tuncali, and Stuart G Silverman. Correlation and simple linear regression. *Radiology*, 227(3):617–628, 2003.

[195] Helene Gelderblom, Machdel Matthee, Marié Hattingh, and Lizette Weilbach. High school learners' continuance intention to use electronic textbooks: a usability study. *Education and Information Technologies*, 24(2):1753–1776, 2019.

[196] Scott B Morris. Estimating effect sizes from pretest-posttest-control group designs. *Organizational research methods*, 11(2):364–386, 2008.

[197] Daren T Nicholson, Colin Chalk, W Robert J Funnell, and Sam J Daniel. Can virtual reality improve anatomy education? a randomised controlled study of a computer-generated three-dimensional anatomical ear model. *Medical education*, 40(11):1081–1087, 2006.

[198] Malinda Carpenter and Michael Tomasello. Joint attention and imitative learning in children, chimpanzees, and enculturated chimpanzees. *Social Development*, 4(3):217–237, 1995.

[199] Masako Hirotani, Manuela Stets, Tricia Striano, and Angela D Friederici. Joint attention helps infants learn new words: event-related potential evidence. *Neuroreport*, 20(6):600–605, 2009.

[200] Juan Garzón, Silvia Baldiris, Jaime Gutiérrez, Juan Pavón, et al. How do pedagogical approaches affect the impact of augmented reality on education? a meta-analysis and research synthesis. *Educational Research Review*, 31:100334, 2020.

[201] Sudha M Srinivasan, Isabel K Park, Linda B Neelly, and Anjana N Bhat. A comparison of the effects of rhythm and robotic interventions on repetitive behaviors and affective states of children with autism spectrum disorder (asd). *Research in Autism Spectrum Disorders*, 18:51–63, 2015.

[202] Michael Rutter, A Bailey, and Catherine Lord. Scq. *The Social Communication Questionnaire. Torrance, CA: Western Psychological Services*, 2003.

[203] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3637–3646, 2017.

[204] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[205] Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *BMVC*, volume 1, page 33. Citeseer, 2010.

[206] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proceedings of 2011 IEEE International Conference on Computer Vision Workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.

[207] Roghayeh Barmaki and Charles E Hughes. Embodiment analytics of practicing teachers in a virtual immersive environment. *Journal of Computer Assisted Learning*, 34(4):387–396, 2018.

[208] Kangsoo Kim, Arjun Nagendran, Jeremy Bailenson, and Greg Welch. Expectancy violations related to a virtual human's joint gaze behavior in real-virtual human interactions. In *Proceedings of the International Conference on Computer Animation and Social Agents*, pages 5–8, 2015.

[209] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9):3046, 2021.

[210] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007.

[211] Juanpablo Andrew Heredia Parillo. An automatic emotion recognition system that uses the human body posture. 2021.

[212] Conner J Black, Abigail L Hogan, Kayla D Smith, and Jane E Roberts. Early behavioral and physiological markers of social anxiety in infants with fragile x syndrome. *Journal of Neurodevelopmental Disorders*, 13(1):1–9, 2021.

[213] Zahid Akhtar and Tiago H Falk. Visual nonverbal behavior analysis: the path forward. *IEEE MultiMedia*, 25(2):47–60, 2017.

[214] Francesca Capozzi and Jelena Ristic. How attention gates social interactions. *Annals of the New York Academy of Sciences*, 1426(1):179–198, 2018.

[215] C Lord, M Rutter, PC DiLavore, S Risi, K Gotham, and SL Bishop. Autism diagnostic observation schedule,(ados-2) modules 1-4. *Los Angeles, California: Western Psychological Services*, 2012.

[216] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[217] Brian Scassellati. How social robots will help us to diagnose, treat, and understand autism. In *Robotics research*, pages 552–563. Springer, 2007.

[218] Xiang Lian, Wilson Cheong Hin Hong, Xiaoshu Xu, Kolletar-Zhu Kimberly, and Zhi Wang. The influence of picture book design on visual attention of children with autism: a pilot study. *International Journal of Developmental Disabilities*, pages 1–11, 2022.

[219] Julia Vacas, Adoración Antolí, Araceli Sánchez-Raya, Carolina Pérez-Dueñas, and Fátima Cuadrado. Visual preference for social vs. non-social images in young children with autism spectrum disorders. an eye tracking study. *Plos one*, 16(6):e0252795, 2021.

[220] Ahmed Hassan, Niels Pinkwart, and Muhammad Shafi. Serious games to improve social and emotional intelligence in children with autism. *Entertainment computing*, 38:100417, 2021.

[221] Phil Wai Shun Leung, Shirley Xin Li, Carmen Sze Oi Tsang, Bellavista Long Ching Chow, and William Chi Wai Wong. Effectiveness of using mobile technology to improve cognitive and social skills among individuals with autism spectrum disorder: Systematic literature review. *JMIR mental health*, 8(9):e20892, 2021.

[222] Louis Tay, Sang Eun Woo, Louis Hickman, Brandon M Booth, and Sidney D'Mello. A conceptual framework for investigating and mitigating machine-learning measurement bias (mlmb) in psychological assessment. *Advances in Methods and Practices in Psychological Science*, 5(1):25152459211061337, 2022.

[223] Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. Early prediction of visitor engagement in science museums with multimodal learning analytics. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 107–116, 2020.

[224] Paulo Blikstein. Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, page 102–106, New York, NY, USA, 2013. Association for Computing Machinery.

Appendix A

IRB/HUMAN SUBJECTS APPROVAL

Institutional Review Board
210H Hullihen Hall
Newark, DE 19716
Phone: 302-831-2137
Fax: 302-831-2828

DATE:	May 13, 2021
TO:	Anjana Bhat, PhD
FROM:	University of Delaware IRB
STUDY TITLE:	[637082-12] Embodied Rhythm Interventions for Children with Autism Spectrum Disorders
SUBMISSION TYPE:	Continuing Review/Progress Report
ACTION:	APPROVED
APPROVAL DATE:	May 13, 2021
EXPIRATION DATE:	May 17, 2022
REVIEW TYPE:	Expedited Review
REVIEW CATEGORY:	Expedited review category # *(9)*

Thank you for your Continuing Review/Progress Report submission to the University of Delaware Institutional Review Board (UD IRB). The UD IRB has reviewed and APPROVED the proposed research and submitted documents via Expedited Review in compliance with the pertinent federal regulations.

As the Principal Investigator for this study, you are responsible for and agree that:

- All research must be conducted in accordance with the protocol and all other study forms as approved in this submission. Any revisions to the approved study procedures or documents must be reviewed and approved by the IRB prior to their implementation. Please use the UD amendment form to request the review of any changes to approved study procedures or documents.
- Informed consent is a process that must allow prospective participants sufficient opportunity to discuss and consider whether to participate. IRB-approved and stamped consent documents must be used when enrolling participants and a written copy shall be given to the person signing the informed consent form.
- Unanticipated problems, serious adverse events involving risk to participants, and all non-compliance issues must be reported to this office in a timely fashion according with the UD requirements for reportable events. All sponsor reporting requirements must also be followed.

Oversight of this study by the UD IRB REQUIRES the submission of a CONTINUING REVIEW seeking the renewal of this IRB approval, which will expire on May 17, 2021. A continuing review/progress report form and up-to-date copies of the protocol form and all other approved study materials must be submitted to the UD IRB at least 45 days prior to the expiration date to allow for the required IRB review of that report.

If you have any questions, please contact the UD IRB Office at (302) 831-2137 or via email at hsrb-research@udel.edu. Please include the study title and reference number in all correspondence with this office.

Institutional Review Board
210H Hullihen Hall
Newark, DE 19716
Phone: 302-831-2137
Fax: 302-831-2828

DATE:	May 5, 2022
TO:	Anjana Bhat, PhD
FROM:	University of Delaware IRB
STUDY TITLE:	[637082-13] Embodied Rhythm Interventions for Children with Autism Spectrum Disorders
SUBMISSION TYPE:	Continuing Review/Progress Report
ACTION:	APPROVED
APPROVAL DATE:	May 5, 2022
EXPIRATION DATE:	May 17, 2023
REVIEW TYPE:	Expedited Review
REVIEW CATEGORY:	Expedited review category # *(9)*

Thank you for your Continuing Review/Progress Report submission to the University of Delaware Institutional Review Board (UD IRB). The UD IRB has reviewed and APPROVED the proposed research and submitted documents via Expedited Review in compliance with the pertinent federal regulations.

As the Principal Investigator for this study, you are responsible for and agree that:

- All research must be conducted in accordance with the protocol and all other study forms as approved in this submission. Any revisions to the approved study procedures or documents must be reviewed and approved by the IRB prior to their implementation. Please use the UD amendment form to request the review of any changes to approved study procedures or documents.
- Informed consent is a process that must allow prospective participants sufficient opportunity to discuss and consider whether to participate. IRB-approved and stamped consent documents must be used when enrolling participants and a written copy shall be given to the person signing the informed consent form.
- Unanticipated problems, serious adverse events involving risk to participants, and all non-compliance issues must be reported to this office in a timely fashion according with the UD requirements for reportable events. All sponsor reporting requirements must also be followed.

Oversight of this study by the UD IRB REQUIRES the submission of a CONTINUING REVIEW seeking the renewal of this IRB approval, which will expire on May 17, 2023. A continuing review/progress report form and up-to-date copies of the protocol form and all other approved study materials must be submitted to the UD IRB at least 45 days prior to the expiration date to allow for the required IRB review of that report.

If you have any questions, please contact the UD IRB Office at (302) 831-2137 or via email at hsrb-research@udel.edu. Please include the study title and reference number in all correspondence with this office.

JOHNS HOPKINS
U N I V E R S I T Y

Homewood Institutional Review Board
3400 N. Charles Street
Baltimore MD 21218-2685
410-516-6580
http://web.jhu.edu/Homewood-IRB/

Michael McCloskey, PhD
Chair

Date: March 31, 2017

PI Name: Nassir Navab
Study #: HIRB00005021
Study Name: Educational Impact of an Augmented Reality Mirror System

Date of Review: 3/31/2017
Date of Approval: 3/31/2017
Expiration Date: 3/31/2020

The above referenced study has been *approved*.

Review Type:	Exempt
Funding Agency:	Other JHU Science of Learning Insitute
Grant or Contract Number:	N/A
International Sites:	No
Maximum number of participants:	350
Vulnerable populations:	JHU Students
Consent process:	Waiver of written consent (Oral Informed Consent)
Assent Process:	Written assent

Please keep in mind that it is your responsibility to inform the HIRB of any adverse consequences to participants that occur in the course of the study, as well as any complaints from participants regarding the research. In conducting this research, you are required to follow the requirements listed in the *HIRB Policies and Procedures Manual*.

Approved Documents:

Oral Consents:
Oral_consent_Script_magic mirror-Treatment.docx
Oral_consent_Script_magic mirror-Control.docx

Recruiting Materials:
Email script

Study Team Members:
Nassir Navab
Gregory Hager
Bernhard Fuerst
Roghayeh Barmaki
Richard Shingles
Rebecca Pearlman
Kevin Yu
Felix Bork

APPROVAL IS GRANTED UNDER THE TERMS OF **FWA00005834** FEDERAL-WIDE ASSURANCE OF COMPLIANCE WITH DHHS REGULATIONS FOR PROTECTION OF HUMAN RESEARCH SUBJECTS

Appendix B

PERMISSIONS

Chapter 3 is a minor reversion of "Collaboration Analysis Using Object Detection." published in Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019), and has been reproduced here with the permission of the copyright holder. The link is https://drive.google.com/file/d/1yznkJQl-bkP1y5sIjRjm8RwaInXNqp-k.

Chapter 4 is a slightly modified version of "Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction." published in Australasian Journal of Educational Technology and has been reproduced here with the permission of the copyright holder https://doi.org/10.14742/ajet.6436.

Chapter 5 is a minor reversion of "An Automated Mutual Gaze Detection Framework for Social Behavior Assessment in Therapy for Children with Autism." published in Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI'21) and has been reproduced here with the permission of the copyright holder https://doi.org/10.1145/3462244.3479882.