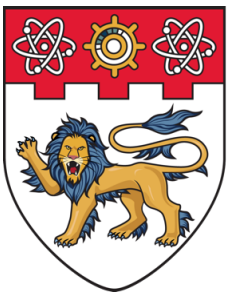

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

NON-VERBAL INFORMATION ESTIMATION IN MULTI-PARTY HUMAN-ROBOT/VIRTUAL HUMAN INTERACTION

ZHANG ZHIJIE

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2023

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other university or institution.

17/01/2023

Date

Zhang Zhijie

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

17/01/2023

.....

Date

.....

Prof. Zheng Jianmin

Authorship Attribution Statement

This thesis contains materials from three papers published in the following peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as **Z. Zhang, J. Zheng and N. Magnenat Thalmann, "Engagement Intention Estimation in Multiparty Human-Robot Interaction," in *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 117–122.**

The contributions of the co-authors are as follows:

- Prof. Jianmin Zheng and Prof. Nadia Magnenat Thalmann provided the initial research directions.
- Prof. Jianmin Zheng and I proposed the model frameworks and developed the new methods.
- I implemented the proposed methods and conducted the experiments.
- I wrote the manuscript drafts. These drafts were revised by Prof. Jianmin Zheng and Prof. Nadia Magnenat Thalmann.

Chapter 4 is published as **Z. Zhang, J. Zheng and N. Magnenat Thalmann, "Engagement estimation of the elderly from wild multiparty human-robot interaction," *Computer Animation and Virtual Worlds (CAVW)*, 2021, pp. e2120.**

The contributions of the co-authors are as follows:

- Prof. Jianmin Zheng and I provided the initial research directions.
- Prof. Nadia Magnenat Thalmann provided the robot platform Nadine.
- Prof. Nadia Magnenat Thalmann provided the dataset to me, which she and her colleagues collected from a nursing home in another project.
- Prof. Jianmin Zheng and I proposed the model frameworks and developed new methods.
- I implemented the frameworks and technical approaches, and conducted experiments.
- I wrote the manuscript drafts. These drafts were revised by Prof. Jianmin Zheng.

Chapter 5 is published as **Z. Zhang, J. Zheng and N. Magnenat Thalmann, "Real and Apparent Personality Prediction in Human-Human Interaction," in *International Conference on Cyberworlds (CW)*, 2022, pp.**

187–194. and **Z. Zhang, J. Zheng and N. Magnenat Thalmann, "Context-Aware Personality Estimation and Emotion Recognition in Social Interaction,"** ***The Visual Computer (TVCJ)*****, 2023.**

The contributions of the co-authors are as follows:

- Prof. Jianmin Zheng provided the initial research directions.
- Prof. Jianmin Zheng and I proposed the model frameworks and developed technical approaches.
- I designed and implemented the methods, and conducted experiments.
- I wrote the manuscript drafts. These drafts were revised by Prof. Jianmin Zheng and Prof. Nadia Magnenat Thalmann.

17/01/2023

Date

Zhang Zhijie

Acknowledgements

I wish to express my greatest gratitude to my supervisor, Prof. Jianmin Zheng, for his invaluable help and patient guidance. My Ph.D. study journey would not have been possible without him, who generously provided his knowledge and advice.

I would like to express my deepest appreciation to my co-supervisor, Prof. Nadia Magnenat Thalmann, for the support she gave to me during my first and second years. Both the guidance in research directions and the provision of the research platforms have been precious and unparalleled opportunities in my research life. In particular, I really appreciate that she shared with me the social robot, virtual agents and some datasets, which she created or collected from her other research projects.

I am also grateful to my colleagues for their selfless help and moral support, which have impacted and inspired me, making these years memorable. Special thanks go to Dr. Li Tian for his encouragement and advice. Special thanks go to Ms. Nidhi Mishra and Ms. Gauri Tulsulkar for their contributions to the experimental data collection.

Finally, I am very grateful to my family and friends for their help and financial support. The trust and encouragement from my parents are the spiritual force for me to keep my determination and motivation during this journal. Many thanks to my friends, Yang Sheng, Mei Liu, and Jiafan Li, who accompanied me. I would also like to thank all the people who have been a part of my life, as it is your kindness that has sustained me through this journey.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Average Pooling
APP	Automatic Personality Perception
APR	Automatic Personality Recognition
APS	Automatic Personality Synthesis
ASD	Autism Spectrum Disorder
ASR	Automatic Speech Recognition
AUs	Action Units
BBX	Bounding Boxes
BFI	Big-Five Inventory
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DFA	Deterministic Finite Automaton
EE	Engagement Estimation
EPWDS	Engagement of a Person with Dementia Scale
ER	Emotion Recognition
ERI	Elderly-Robot Interaction
FC	Fully Connected Layer
FOVA	Focus of Visual Attention
FPS	Frame per Second
GAL	Graph Attention Layer
GAT	Graph Attention Network
GNN	Graph Neural Network
HHI	Human-Human Interaction

HRI	Human-Robot Interaction
I2P	Integrated Interaction Platform
IA	Intelligent Agent
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NAS	Neural Architecture Search
NB	Naive Bayes
NLP	Natural Language Processing
NVC	Non-Verbal Communication
PAD	Pleasure-Arousal-Dominance
PE	Personality Estimation
RF	Random Forest
RL	Reinforcement Learning
SSP	Social Signal Processing
SVM	Support Vector Machine
VA	Valence-Arousal
VNN	Vanilla Neural Network

Abstract

Robots and virtual agents have been deployed in various fields, and are playing an increasingly important role in human being's daily lives. Thus, these intelligent agents (IAs) are required to interact with their users appropriately. To achieve this goal, IAs need to understand human social signals, before generating socially acceptable responses. However, current multi-party social human-robot interaction (HRI) is still far from being satisfactory. Unlike dyadic HRI, multi-party HRI involves more than one participant in the interaction, so with the increase in the participant number, IAs face more challenging tasks. The overall objective of this research is to investigate and develop new techniques to empower robots or virtual agents with the ability to understand the behaviors, intentions, and affects of the participants in multi-party social interaction, which helps the agent manage multi-party issues in social HRI. Specifically, this thesis presents new methods to analyze and estimate four types of non-verbal social information in multi-party human-robot interaction scenarios, namely (i) engagement intention estimation, (ii) engagement estimation during interaction, (iii) personality estimation, and (iv) emotion recognition.

To understand the dynamics in multi-party social HRI, estimating user engagement intention is chosen as the first task to be studied. A method to estimate if people have willingness to join in a conversation in multi-party scenarios is first presented, which helps endow IAs with the capability of detecting potential participants. The method is built on the deep neural networks, which takes image features and social signals as input, making use of general information conveyed in images, semantic social cues proven by social psychology studies, and temporal information in the sequence of inputs. The network is designed to have a multi-branch structure with the flexibility of accommodating different types of inputs. Also, the signal transition in multi-party human-robot interaction scenarios is discussed. The method

is evaluated on three datasets with social signals and/or images as inputs. The results show that the proposed method can infer human engagement intention well.

When an interaction starts, the goal of engagement estimation (EE) is to infer the inner state of a participant attributing to being together with the other participants and continuing the interaction. Moreover, the use of social robots in healthcare systems or nursing homes to assist the elderly and their caregivers will be becoming common. It is proposed that a supervised machine learning method to estimate the engagement state of the elderly in a multi-party human-robot interaction scenario from real-world video recordings. The method is built upon the basic concept of engagement in geriatric psychiatry and HRI video representations. It adapts pre-trained models to extract behavior, affective, and visual signals to form multi-modal features. These features are then fed into a neural network made of a self-attention mechanism and average pooling for individual learning, a graph attention network for group learning, and a fully connected layer to estimate engagement. The proposed method is evaluated using 43 wild multi-party elderly robot interaction (ERI) videos. The experimental results show that the proposed method is capable of detecting the key participants and estimating the engagement state of the elderly effectively. Also, the signals from side participants in the main interaction group considerably contribute to the EE of the elderly in the multi-party ERI.

Finally, personality and emotion are investigated in multi-party social interaction, which have great influences on people's cognition and behavior. There are fewer works on personality estimation (PE). The contribution of personality to emotion recognition (ER) has also not been adequately studied. Therefore, a context-aware deep learning framework has been proposed, which automatically estimates the personality of a target person in human-human social interaction scenarios, based on the target person's own and the interlocutor's body behavioral and facial information. Then, this network is expanded to form a context-aware and personality-based emotion recognition framework, *i.e.*, first estimating the personality and then recognizing the emotions based on the estimated personality. A set of experiments have been conducted showing that the proposed method has good performance in both personality and emotion experiments.

Chapter 1

Introduction

1.1 Background

Robots and virtual agents have been developed very rapidly in the past decades. They have left their professional fields and factories, and have been deployed in various areas such as the service industry [10], skill training [11, 12], education [13], and entertainment industries [14], taking roles of robot services, tutors, tourism, and even companion characters. These applications are different from traditional industrial robots because intelligent agents (IAs) are required to interact with their users by simulating human functions and behaving like a human. Based on the tasks that IAs aim to address, they can be categorized into task-oriented systems with a specific goal and non-task-oriented systems that serve in free interactions [15, 16]. However, both of them need to interact socially with their users to a greater or lesser extent, so the social ability to address human-robot interaction (HRI) problems is indispensable.

HRI includes both the creation of robots that assist humans with certain duties or tasks, and the examination of how people and robots interact with each other in this situation [17]. In order to allow users to gain and maintain a good experience in HRI, social agents must be endowed with social intelligence. Conventional research is more task-oriented but ignores the importance of natural and friendly interaction, but socially intelligent agents are getting more and more attention now [18].

Giving robots the ability to understand human behavior and infer implicit intentions can help robots interact with humans more naturally and efficiently. Or in other words, IAs are expected to be able to understand human behaviors, intentions, and affects, and then generate socially acceptable responses based on the analysis and understanding of the participants' social signals. More importantly, this social intelligence can act as a transition smoother or a rapport builder for complex tasks that require user cooperation.

Multi-party social HRI is a face-to-face interaction, in which more than one human participants interact with a social robot, usually forming a small group freely in some ways. Participants can unrestrictedly choose their own positions, body postures, actions, expressions, and even attending or leaving, *etc.*, that is, a face-to-face conversation among multiple people in a free way that is common in daily life. Conventional work on socially interactive robots and virtual agents mainly focuses on one-to-one interactions, *i.e.*, a social robot interacting with a user. However, multi-party interaction involves multiple users, which requires further investigation. Giving socially intelligent agents the ability to participate in one or multiple conversations with more than one participant poses many challenges for designing appropriate visual and language management systems and modules.

1.2 Motivations, Objectives, and Challenges

Social robots have traditionally been designed for dyadic conversations, but they lack the ability to communicate with multiple users in one or multiple concurrent conversations. Social interactions do not always follow a one-to-one conversational situation in daily life, *i.e.*, conversations among human beings often involve more than two participants, leading to multi-party interactions where individuals naturally form groups and communicate with each other. Furthermore, in these conversations, participants are free to join, leave, pause, or resume the conversation at any time, and communication among participants rarely encounters conflicts or interruptions.

Although a large amount of work has attempted to propose solutions to improve robot performance in HRI, reresearch on multi-party social HRI is still very limited. This can be attributed to two primary factors. Firstly, the development and implementation of multi-party interaction processing modules heavily depend on the robust and efficient analysis of dyadic interactions. Meanwhile, as the number of participants increases, social robots are confronted with increasingly demanding tasks compared to dyadic interactions.

Additionally, non-verbal communication serves as a complementary mode of communication that can augment and enrich the conveyed message when combined with speech. It is widely acknowledged that individuals heavily rely on non-verbal cues to express themselves and interpret others' intentions. In computer science, a lot of studies have focused on inferring people's intentions and inner states by analyzing and predicting their non-verbal information.

This thesis hopes to bring the results of psychological and social behavior science into computer science, forming a multidisciplinary perspective to help us design multi-party socially intelligent agents. In order to attain this goal, IAs first should be able to comprehend social cues from multi-party social interactions, and then generate socially acceptable responses. This thesis will focus on the first part, *i.e.*, the analysis and processing of the non-verbal social signals for multiple participants in multi-party interactions. Consequently, the overall objective of this thesis is to empower IAs with the ability to understand the behaviors, intentions, and affects of the participants in multi-party social interaction, which helps the agent manage multi-party issues in social HRI. Specifically, this thesis analyzes non-verbal social information in multi-party human-robot interaction scenarios in four areas: (i) engagement intention estimation, (ii) engagement estimation during interaction, (iii) personality estimation, and (iv) emotion recognition. These non-verbal social signals are anticipated to be utilized in future research to facilitate the generation of more natural and human-like responses that align better with people's expectations.

The prediction of engagement intention, which stands for estimating whether people have the intention to initiate or join a conversation with a robot, is the first task that will be discussed. This function is not only useful in the initial period of

the HRI, but also helpful for inviting potential participants into an existing conversation to form a multi-party interaction. In addition, it can further be extended as a function to manage dynamics, *i.e.*, join and leave problems in multi-party social HRI. Then, once a multi-party interaction begins, the estimation of engagement during this interaction becomes a function that IAs required to have. By estimating the multi-participants' engagement, IAs are able to keep a more balanced multi-party conversation, thus avoiding one dominant participant leading the entire conversation, which is important in collaborative tasks. More importantly, if a particular participant should be given more attention, such as a patient in the medical or psychological treatments, then the robot can change its behavior based on the estimated engagement level, thus keeping the patient's engagement at a high level to improve the treatment outcome. Finally, the users' emotions during HRI are a very important part that cannot be ignored. A robot that just performs tasks cannot be called a socially intelligent robot. IAs can act and speak if they can recognize the emotions of each participant in a multi-party interaction, where emotions are not only mutual between two people, but become a kind of convergence between groups, or confrontation between more subdivided subgroups. Thus, only after sufficient exploration, can robots more comprehensively understand the events occurring in the interaction and further generate their own emotions.

It is not easy to fully accomplish the above tasks. The difficulties of estimating the user engagement intention are that the user's behavior and signals that express intention are non-verbal information that is not easy to recognize, and the time from a user has the intention to actual interaction is of short duration. Existing work extracts relative features for each frame of data and then predicts user engagement intention from an independent time point without adequate exploration of the temporal information. Besides, multi-party interaction is far from being thoroughly studied and always be overlooked. Tackling the above issues will enable HRI to better meet natural interaction between human beings and robots. In terms of engagement estimation during an interaction, many methods have been developed in various scenarios such as classroom or distance learning and healthcare. Conventional approaches use non-verbal cues such as proxemics, body pose, gaze patterns, facial expressions, and context information to build classifiers. Deep

learning approaches have also been developed. However, most previous work assumes the interaction is in a laboratory environment or a dyadic situation. To analyze wild multi-party interaction, understanding the dynamics and stability of the interaction becomes more complicated. Moreover, in unconstrained wild space, moving people, bad lighting, confusing objects, *etc.* make it difficult to interpret the complex environment. Finally, both personality estimation and emotion recognition are the analyses of visual, auditory, and semantic information, which means that existing solutions for these two tasks are intuitively very similar, but are often treated as two different tasks in research. Although there are some works trying to estimate real or apparent personality, current automatic vision-based or multimodal-based personality prediction methods do not perform well. There are many approaches to emotion recognition. Some group or scenario-based emotion analysis methods are introduced, but they all deal with them as a whole and do not take into account the interactions of participants in the social interaction, which does not correspond to reality, *i.e.*, people's expressions and actions not only depend on their personality, but also are influenced by the external environment and the behaviors of the interlocutors. This interaction between participants leads to the transfer of emotions, which is difficult to capture when such rapid and subtle changes occur.

To this end, this thesis tries to provide some solutions to the above-mentioned problems and challenges.

- **Engagement intention estimation before the interaction starts.** Engagement intention prediction in a multi-party scenario is studied. A method to estimate if people have willing to join in a conversation in multi-party scenarios is first presented, which helps to endow IAs with the capability of detecting potential participants. The method is built on convolutional neural network (CNN) and long short-term memory (LSTM), which takes image features and social signals as input, making use of general information conveyed in images, semantic social cues proven by social psychology studies, and temporal information in the sequence of inputs. The network is designed to have a multi-branch structure with the flexibility of accommodating different types of inputs. Also, the signal transition in multi-party human-robot interaction scenarios is discussed. The method is evaluated on three datasets with

social signals and/or images as inputs. The results show that the proposed method can infer human engagement intention well.

- **Engagement estimation during interaction.** When an interaction starts, the goal of engagement estimation (EE) is to infer the inner state of a participant attributing to being together with the other participants and continuing the interaction. Moreover, the use of social robots in healthcare systems or nursing homes to assist the elderly and their caregivers will be becoming common. It is proposed that a supervised machine learning method to estimate the engagement state of the elderly in a multi-party human-robot interaction scenario from real-world video recordings. The method is built upon the basic concept of engagement in geriatric psychiatry and HRI video representations. It adapts pre-trained models to extract behavior, affective, and visual signals to form multi-modal features. These features are then fed into a neural network made of a self-attention mechanism and average pooling for individual learning, a graph attention network for group learning, and a fully connected layer to estimate engagement. The proposed method is evaluated using 43 wild multi-party elderly robot interaction (ERI) videos. The experimental results show that the proposed method is capable of detecting the key participants and estimating the engagement state of the elderly effectively. Also, the signals from side participants in the main interaction group considerably contribute to the EE of the elderly in the multi-party ERI.

- **Personality estimation and emotion recognition.** Finally, personality and emotion, investigated in multi-party social interaction, have great influences on people's cognition and behavior. There are fewer works on personality estimation (PE). The contribution of personality to emotion recognition (ER) has also not been adequately studied. Therefore, a context-aware deep learning framework has been proposed, which automatically estimates the personality of a target person in human-human social interaction scenarios, based on the target person's own and the interlocutor's body behavioral and facial information. Then, this network is expanded to form a context-aware and personality-based emotion recognition framework, *i.e.*, first estimating the personality and then recognizing the emotions based on the estimated personality. A set of experiments have been conducted showing that the

proposed method has good performance in both personality and emotion experiments.

1.3 Contributions

The primary contribution of this thesis is to model the perceptual and affective processes of socially intelligent agents such as robots and virtual human. This is achieved through the analysis of non-verbal social signals from multi-party social interactions, particularly in human-robot social interactions. The behaviors, intentions, and affective states of the participants in these interactions are estimated, which is crucial for the development of social agents. These evaluated social signals and affective states are expected to be integrated into social robot platforms, optimizing the generation of natural and human-like responses during multi-party conversations. Specifically, according to different tasks, the contributions and novelties of this thesis can be stated as follows.

- **Engagement intention estimation before multi-party social HRI**
  - A novel architecture is designed to estimate the engagement intention of potential participants in multi-party HRI scenarios. This new architecture is characterized by its multi-branch and adaptable input structures.
  - The findings from psychological research are reviewed and summarized to identify high-level social signals that can be combined with image features to serve as inputs.
  - A new neural network model based on CNN and LSTM is proposed to estimate the engagement intention of potential participants.
  - A novel feature transition method is designed to interpret multi-party social signals, which is essential to enable robots to perform well in such scenarios.
  - The experimental results indicate that the proposed approach has good performance in terms of accurately estimating engagement intention by utilizing multi-modal features.

- **Engagement estimation during multi-party social HRI**
  - A novel and automatic method for estimating participants' engagement levels in multi-party HRI is proposed.
  - More challenging and less explored wild multi-party interactions between elderly people and a robot are analyzed, compared to existing approaches.
  - A novel engagement estimation framework for such scenarios is designed by combining the engagement studies from psychiatry with computer vision techniques, where behavioral, affective, and visual engagement and their features are investigated.
  - A new deep learning model is constructed, which includes self-attention networks and graph attention networks to learn individual and group information, which efficiently improves the performance of engagement estimation.
  - A new dataset is created for studying engagement in multi-party elderly-robot interaction in the natural environment. The dataset includes multi-view video recordings and labeled annotations.
- **Personality estimation and emotion recognition in social interaction**
  - A novel approach is proposed to jointly estimate personality and recognize the emotions of participants in social interaction scenarios.
  - Different from the prior art, the newly proposed method employs the same architecture to analyze personality and emotions, which in addition is capable of estimating both apparent and real personality.
  - The social interaction context is emphasized and utilized. The multimodal data from the target individual and interlocutor(s) are utilized, forming a context-aware structure to estimate personality, and then use personality to improve the accuracy of emotion recognition.
  - A set of experiments have been conducted to examine the impact of information from the target individual and interlocutor(s) on personality estimation, as well as the effects of personality on emotion recognition.

1.4 Outline of the Thesis

This thesis is organized as follows.

- **Chapter 1** introduces the research background, motivations and objectives, and major contributions of this thesis.
- **Chapter 2** reviews the state-of-the-art, relative techniques and supporting materials for multi-party social HRI. In particular, an overview of the development in HRI is discussed first, including basic HRI, social HRI, and multi-party social HRI. Then, the relevant concepts of social interaction are discussed from the perspective of interdisciplinary studies, after which the related work in user engagement intention estimation, engagement estimation, personality estimation, and emotion recognition are reviewed and summarized.
- In **Chapter 3**, a multi-branch CNN-LSTM-based network for engagement intention prediction in a multi-party social HRI scenario is proposed. Human non-verbal cues and features used in the estimation are discussed in this section, after which the signal transition in multi-party human-robot interaction scenarios is discussed.
- **Chapter 4** extends the engagement estimation to the period of interaction ongoing. A supervised deep learning architecture to estimate the engagement state of the elderly in a multi-party social HRI is presented, based on the concepts of engagement in geriatric psychiatry and HRI video representations. It extracts behavioral, affective, and visual signals to form the multi-modal features and then learns individual and group representations from multi-party participants.
- **Chapter 5** investigates personality and emotions in multi-party social interaction. A context-aware deep learning framework has been proposed to estimate the personality of a target person, based on the target person's own and the interlocutor's body behavioral and facial information. An expanded model, context-aware and personality-based emotion recognition framework, is explained for recognizing emotions.

- Finally, **Chapter 6** summarizes the key conclusions of the thesis and provides some future research recommendations on further development and improvement of the multi-party human-robot social interaction.

Chapter 2

Literature Review

This chapter first provides an overview of HRI and recent research into multi-party social HRI in Section 2.1. Thereafter, as this thesis involves many findings and theories of psychology, cognitive and behavioral science, and geriatric-related studies, Section 2.2 reviews the relevant interdisciplinary literature. Then, the state-of-the-art work of user engagement intention estimation, engagement estimation during the interaction, personality estimation, and emotion recognition are reviewed in Section 2.3, Section 2.4, Section 2.5, Section 2.6, respectively. A summary of related work is provided in Section 2.7, which highlights the potential areas for further research.

2.1 Overview of Human-Robot Interaction

2.1.1 Human-Robot Interaction

The study of interactions between humans and robots, or more generally, interactions between humans and artificial agents, is known as human-robot interaction (HRI). Goodrich and Schultz provided a definition of HRI that aims to comprehend, develop, and evaluate robots that are used by or with humans [19]. Dautenhahn placed more stringent requirements on HRI. He believes that HRI needs to study the reactions and attitudes of human beings towards robots with different physical,

technical, and interactive characteristics, so as to design robots that are efficient, widely accepted, satisfy users' emotional needs, and have universal values [20]. Traditionally, because robots are still used by humans as tools, HRI is inherently present in all robotics.

The interaction between humans and robots is affected by the attributes and characteristics of robots, such as the appearance of the robot, the level of autonomy, the way information exchanging, the structure and task of the intended interaction, *etc*. Industrial robots, ground robots, pet robots, and humanoids have diverse appearances, which meet specific needs. Autonomy refers to the extent to which a robot can act automatically, *i.e*., expressed in a spectrum, from being completely controlled by humans to completely having their own decisions, autonomy gradually increases. Then, information is exchanged in visual displays, gestures, speech and natural language, and so on. Finally, HRI also is influenced by the structure and task of the interaction, or in other words, the team structure and the goals to be accomplished.

After understanding the influencing characteristics, the subsequent inquiry pertains to the criteria by which the robot ought to be designed in order to fulfill specific requirements. Dautenhahn emphasized three directions: robot-centered HRI, human-centered HRI, and robot cognition-centered HRI [18]. The robot-centered HRI considers robots as autonomous individual that achieves goals according to their motivations and complete specific tasks by interacting with humans. Human-centered HRI focuses on helping robots accomplish tasks in a way that is acceptable to humans. It explores users' experiences and opinions of a robot's appearance and behavior without considering the robot's decision-making process. Finding an appearance, generating behaviors, estimating the intention and requirements of single or multiple users, *etc*., are some typical research problems. Robot cognition-centered HRI places a strong emphasis on the idea that a robot is an intelligent agent. To achieve this objective, researchers have to develop cognition-and-affective processes modules using machine learning techniques. However, these directions often intersect with each other. He also points out that, by defining socially acceptable behavior, it is possible to design robots that can satisfy users' preferences.

2.1.2 Human-Robot Social Interaction and Social Robots

Social interaction is an important way for us to communicate with others, and it is also the process by which individuals know and shape themselves. Social interaction plays an important role in human life. People exchange information through perception, expression, and decision-making, and at the same time, friendship, empathy, and love, or confrontation, disgust, and jealousy, are also products of social interaction. These abstract concepts are the basis of human society. People more or less have the need to express themselves and socialize with others. Although some people are introverted and some suffer from social disorders, social interaction is inevitable. It happens in daily life anytime and anywhere, and it also has a positive impact on people's mental health.

As more and more robots appear in daily life, people's requirements for HRI are becoming more and more stringent. Compared with the early days, when robots are designed to complete tasks alone guided by a command from humans, or to achieve a goal through interaction with humans, until now, people expect robots to meet their social needs, that is, behave and express like a human in interactions. Research has shown that HRI is influenced by the past interaction experience with humans, and the way people communicate with robots is similar to the way they interact with humans [21], which also explains the expectations people place on robots. Therefore, designing social robots, or robots with social skills and abilities, has become a very important research topic.

The term *social robot* is used in a wide variety of ways. Dautenhahn analyzed the notions of social robots but found the terminologies and degree of robot social intelligence is various [18], *e.g.*, socially situated, sociable, socially intelligent, and socially interactive robots are all given depends on certain research emphasis. Combining the definition from [22–24], Yan *et al.* identified a shared characteristic of social robots. They stated that in order for a robot to qualify as a social robot, it must be able to carry out specific duties and interact with people by following particular social cues and conventions [25]. Moreover, interaction capability is the most important factor for a social robot.

Although a universal definition of a social robot does not exist, it is not disputed that a social robot needs to be socially accepted by human society, so the verbal

FIGURE 2.1: Examples of social robots in different application environments: (a) iCat robot plays chess with children in primary school [1]; (b) Pepper robot teaches children to learn English words [2]; (c) Social robot facilitate home-based intervention for children with ASD [3]; (d) Another social robot for children with ASD [4]; (e) Therapeutic seal robot is used in dementia treatment [5]; and (f) Humanoid robot Nadine interacts with old people in a nursing home [6].

and non-verbal modules that deal with interactions become the core of the study of human-robot interaction. Therefore, socially accepted HRI presents many challenges to artificial intelligence (AI) techniques. In order to design a social robot conforming to social norms and satisfying user expectations, it needs to simulate the perception, decision-making, and behavior generation processes [26]. The evaluation of real situations is based on learnable and evolvable criteria, which are used

to make decisions along with goals. The robot's internal state, such as emotions and goals, will be updated and finally generates natural and emotional language and behaviors.

For the development of social robots, researchers have conducted a lot of studies and experiments. Social robots also play their roles to help humans in different application scenarios, such as education [27], health care [28], public service [29], *etc*. Figure 2.1 provides some examples of social robots situated in different application environments. In the two pictures of the first row, the robot is designed as a teaching assistant, playing chess with children [1] or teaching English words to children [2]. The second row is the applications of social robots in the treatment of children with autism spectrum disorder (ASD) (left: [3] and right: [4]). For the last row of images, on the left, the shown robot is a widely used seal-shaped social robot, which is used in the treatment of dementia. On the right, it is a humanoid robot Nadine, chatting and playing games with the elderly in a nursing home [6]. It is noteworthy that the social robot Nadine is a humanoid social robot, which is more challenging to develop as people may have higher expectations, *e.g.*, human-like behavior, natural language, social skills, *etc*., due to their human-like appearance.

2.1.3 Multi-Party Human-Robot Social Interaction

Multi-party social HRI is a face-to-face interaction, in which more than one human participants interact with a social robot, usually forming a small group freely in some ways. Participants can unrestrictedly choose their own positions, body postures, actions, expressions, and even attending or leaving, *etc*., that is, a face-to-face conversation among multiple people in a free way that is common in daily life. Conventional work on socially interactive robots and virtual agents mainly focuses on dyadic interactions. However, multi-party interaction requires further investigation [30, 31], and studies on analyzing multi-party social HRI in wild and dynamic environments are only decade history [32–34].

Bohus and Horvitz [35] explain the challenges in multi-party HRI. As shown in Figure 2.2, the possible structures of HRI can be categorized into four groups based

on the number of participants, the number of interactions, and dynamics. The intelligent agent dynamically handles situations of multi-participant interaction, which involves two or more participants, and this number may vary. The join of the new participants and the leave of existing participants are allowed at any point in time [36]. The robot is actively engaged in at most one interaction but it can simultaneously keep track of additional, suspended interactions. The ability to work with several persons is a skill that IAs that provide services in the open world need, *i.e.*, the abilities to detect, track, sense, comprehend people over time, and to infer their goals, needs, and attention.

Designing an HRI agent involves addressing many issues, particularly in the context of multi-party social HRI scenarios. These scenarios involve a larger number of complex tasks compared to one-to-one scenarios. Some of these tasks are unique to multi-party interactions, while others differ significantly from their one-to-one counterparts. The following paragraphs delve into the tasks associated with multi-party social HRI and draw a comparison between dyadic and multi-party interactions.

- **Engagement intention estimation before interaction**. In two-party interaction, the robot system only needs to infer and track a single participant's engagement state, where all social signals expressed by the participant directly to the intelligent agent [36–41]. However, in multi-party interaction, not only do more than one participant's social cues have to be detected, but the signals may direct to the other participants rather than the robot. Therefore, a conversion of signals has to be conducted in the data processing stage.

- **Engagement estimation during interaction**. During HRI, if IAs can recognize the engagement state of the participant, it helps the IAs respond properly to maintain long-term interaction or to produce appropriate social behavior for the people to feel a sense of belonging. Compared to dyadic HRI, the results of estimating the engagement of multiple people enable the robot to make decisions, so that balance the conversation to avoid some people being left out.

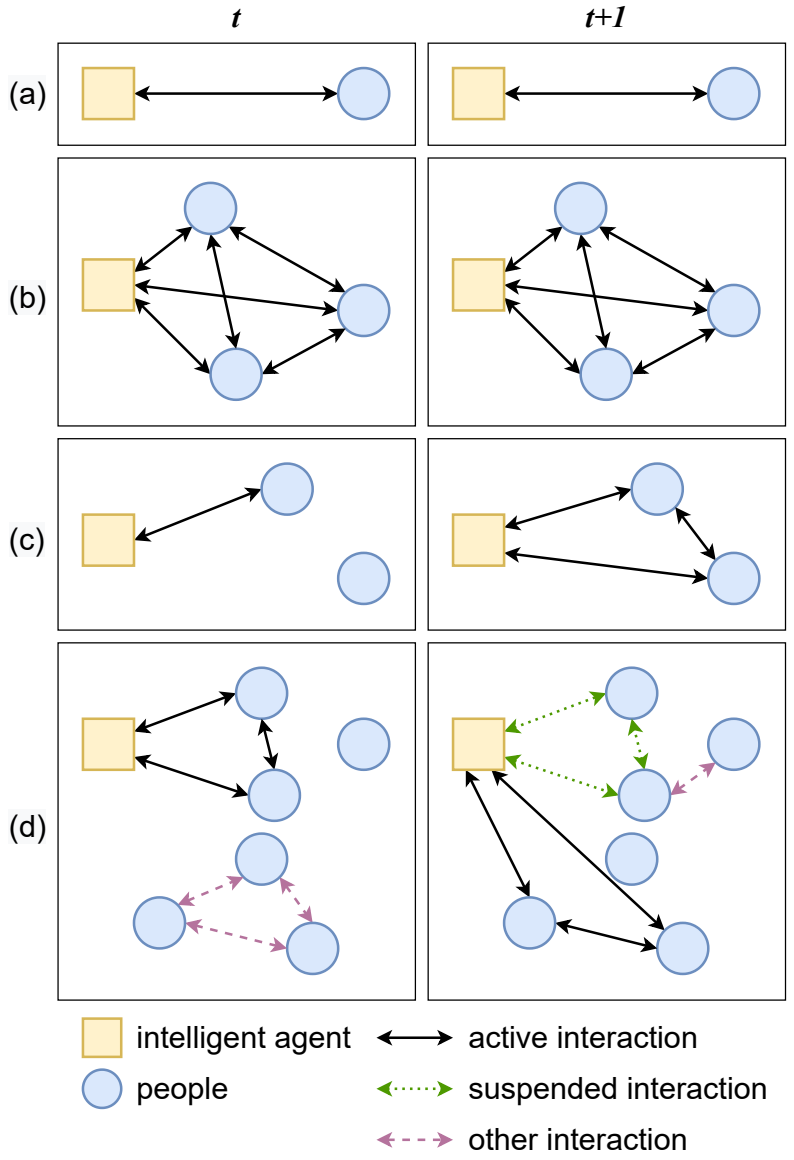

FIGURE 2.2: Illustration of four human-robot interaction structures: (a) single-party HRI; (b) fixed multi-participant HRI; (c) open multi-participant HRI; and (d) open multi-party HRI.

- **Emotion recognition in multi-party interaction**. Generally speaking, emotion recognition is to estimate a person's current emotional state through a person's social signals, which can be linguistic semantic analysis or visual non-verbal signal processing. Traditional visual methods use facial expressions and body postures for inference, but ignore the context information in multi-person social situations, such as group-level emotions, models of emotion generation and change, emotional confrontation and convergence in

groups, and so on. The emotion recognition of multiple people considers the above factors, so as to estimate more accurate emotion states.

- **Conversational roles identification**. In dyadic social HRI, the robot and participant task roles of speaker or addressee, whereas in multi-party social HRI, many new roles appear such as side participants, bystanders, overhearers, and passersby [42–44]. Therefore, it is important to clarify the role of each person captured by the camera. Side participants are the people who are listening to the current conversation and might join in and speak in the future. Bystanders and overhearers are those who listen to the conversation but do not participate. Passersby are the less important people and just pass in the background. Specifically, effectively assigning speech to associated speakers was the main challenge in speaker diarization, which involves allocating speech signals to participants in a dialogue [45–47]. To identify human speakers, audio-visual features like the location of participants, facial landmarks, eye gaze, and sound localization might be used [48]. Compared to speaker diarization, addressee recognition is an opposite task that identifies the addressee in a conversation. The objective of this task is to avoid overlap or responses to unrelated messages. A participant may speak to the other one with the name or can use non-verbal cues such as gaze and pointing gestures to hint addressee [43, 49].

- **Turn-taking management**. People in a conversation understand who is signaling what to whom, so they leverage this understanding to minimize overlaps and conflicts. Robots also need to be endowed with the ability to handle turn-taking [50]. In two-party interaction [51, 52], even in the cases that the dialogue is mixed-initiative, the turn-taking model is simpler whereas participants need to compete for the turn in multi-party conversation [32, 53, 54].

Some research groups also develop an IA system that is, to some extent, able to manage a multi-party conversation in different scenarios. Bohus and Horvitz design a socially situated dialogue system in open environments where two users interact with a virtual human face [50]. Their system included face and pose tracking and estimation of the focus of visual attention (FOVA). They also use this

system to develop an engagement intention detection and turn-taking management model, where participants can initiate, terminate, or join a conversation in an open environment. In the JAMES project, Foster *et al.* develop a robotic bartender that can serve drinks to multiple customers [55], after that they improve the bartender with the engagement prediction model to smooth the service [40]. Kondo *et al.* develop a system for multi-party interaction with a female humanoid who can generate a multi-party gesture [56]. The system conversation is allowed to be interrupted and then the robot can make smooth transitions from a current gesture to the next one. Yumak *et al.* propose a multi-party robot system with tracking and fusion components including speaker diarization and addressee detection [31]. Moreover, their work includes a case study involving a human-robot-virtual agent interaction, which allows the information to transition from a robot to a virtual agent.

2.2 Interdisciplinary Study of Social Interaction

HRI examines the social behavior of humans and the communication between users and robots. Such work is intrinsically interdisciplinary and calls for contributions from various areas, including computer science, robotics, psychology, cognitive and behavioral science, human factors and ergonomics, and other fields. This section will provide reviews from interdisciplinary perspectives. Specifically, this thesis uses non-verbal information, mainly visual signals, to analyze multi-party social interactions, so non-verbal communication (NVC) and social signal processing (SSP) are discussed in Section 2.2.1, after which engagement-related concepts are reviewed in Section 2.2.2. Personality and emotion theories and their applications are presented in Section 2.2.3 and Section 2.2.4 respectively.

2.2.1 Non-Verbal Communication and Social Signal Processing

Non-Verbal Communication. In general, non-verbal communication is employed as a second form of communication that can be used in conjunction with

speech to alter and enrich what is being conveyed [57]. It is the unspoken dialogue that creates shared meaning in social interactions. There is ample evidence showing that people rely substantially on non-verbal cues to express themselves and interpret others' intentions. According to research, human beings prefer to believe non-verbal information rather than verbal information when they conflict and rely on non-verbal messages to assess the attitudes and feelings of others [58].

Non-verbal communication can be categorized into kinesics, proxemics, haptics, chronemics, vocalics, and presentation [59]. Only kinesics and proxemics are taken into consideration in this thesis. Kinesics is the study of non-verbal communication, which includes body language, positioning, facial expressions, gestures, *etc.*, which conveys rich interpersonal, social, and contextual information. The research tasks in kinesics include hand gestures, body movements, eye gaze, and facial expressions. Proxemics deals with how humans use space and its effects on human behavior, communication, and social interaction. Hall categorized the social space into public, social, personal, and intimate based on the distance in human social interactions [60].

The generation of human non-verbal behavior is influenced by many factors, *e.g.*, personal determinants, conscious or unconscious goals, environments, perceptual process, and cognitive-affective processes [58, 61, 62]. Personal determinants include biology, culture, gender, and personality; perceptual and cognitive-affective information are the outcomes of processing interaction and interlocutor(s). Taking dyadic interaction as an example (Figure 2.3), two persons form an interaction and communicate through non-verbal behaviors. Background setting and personal determinants are the high-level factors, where determinants influence the decision-making, perceptual and cognitive-affective processes. At the same time, the latter two interact with each other and influence a person's behaviors together with the determinants.

Social Signal Processing. Social signal processing aims to recognize human behavior and interpret it. Psychologists have studied the processing mechanism of social signals in human brains and found that extracting social information such as facial expression, body language, and tone of voice from communication is often unconscious [63, 64]. In addition, as mentioned before, the foundation of social

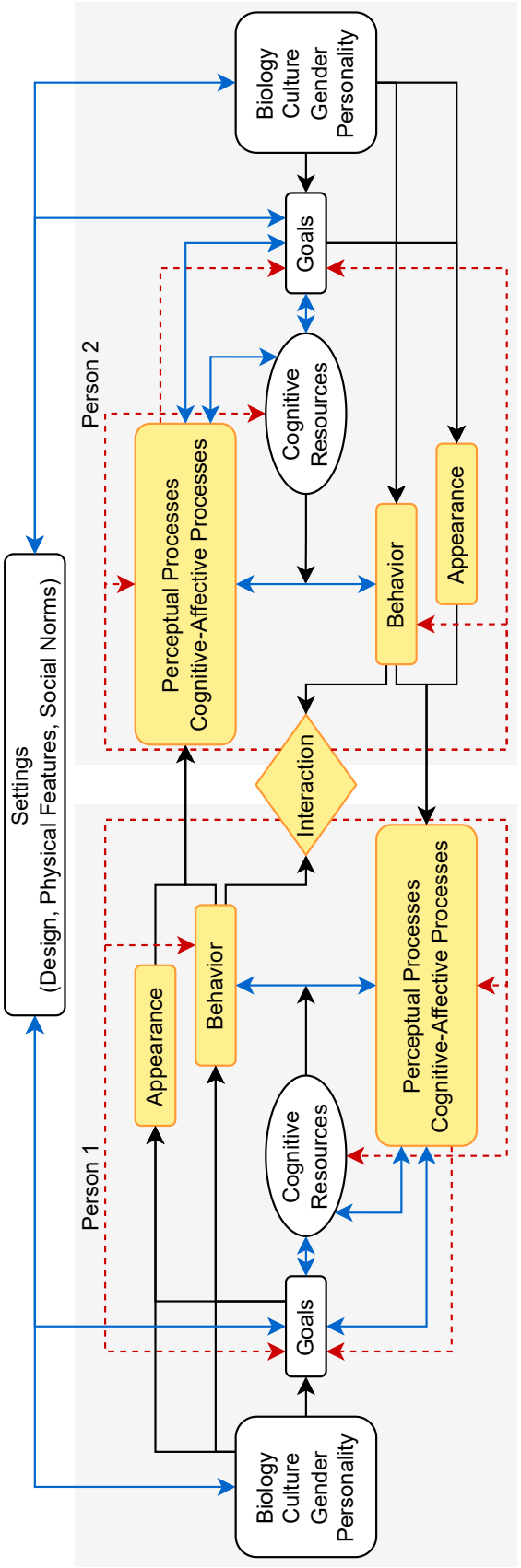

FIGURE 2.3: Systematic model of dyadic non-verbal communication.

intelligence is the capacity to recognize and manage social cues from those with whom IAs are interacting. Therefore, giving socially intelligent robots the ability to understand human behavior through social signal processing is the core of making them intelligent individuals.

In contrast to the semantic and sentiment analysis in natural language processing tasks, the research tasks of social signal processing focus on the interpretation of human non-verbal behaviors displayed by people during social interactions. More specifically, the non-verbal behaviors here include gesture, gaze, facial expression, proxemics, posture, behavioral mirroring, *etc*.

Alessandro *et al*. distinguish between *non-verbal behavior* and *non-verbal cues*, arguing that non-verbal behavior is a continuous source of signals that convey information about people's feelings, psychological states, personality, and other characteristics through a wide range of non-verbal behavior cues, which are mostly perceived and displayed unconsciously [65]. The term non-verbal behavioral cues are used to describe changes in muscle movement or physiological activity over time, usually in short intervals, from milliseconds to minutes, and this is what distinguishes it from non-verbal behavior (minutes to hours). In the computer science area, however, researchers do not distinguish between these two concepts. Most of the work treats both as the same concept, so social behaviors, social signals, and social cues refer to the same concept, in the remaining discussion.

In order to use social signals in socially intelligent agents, developers must first define social signals in terms of their characteristics and then develop automatic detection methods. However, human non-verbal behaviors are difficult to detect; and even when they can be measured, attempting to infer and interpret them is still quite challenging [66]. In order to solve these difficulties, this thesis uses computer vision techniques to analyze these social behaviors via deep learning approaches.

2.2.2 Engagement: Definition, Elements, and Applications

This section reviews related work on engagement, particularly the definition and scope of engagement, clarifications of the relationship and focuses on engagement

intention and engagement during an interaction, and applications and significance of engagement in human-robot social interaction, *e.g.*, in geriatric psychiatry.

Engagement intention. Many studies have proved that non-verbal signals, for example, gestures and postures, face and eye behavior, and space and environment, convey rich information about engagement [63, 67, 68]. Concretely, proxemic features are commonly adopted social cues in HRI [60, 69, 70]. Hall *et al.* emphasize the impact of the use of space on interpersonal communication and introduce a spatial model that horizontally categories the surrounding area of a person into intimate, personal, social, and public spaces (from the near space to the far space) based on different levels of social activities on interpersonal communication [60, 69]. Hall's study of proxemics is valuable in evaluating the way people interact with others in daily life [37]. Kendon and Ferber [71] identify six stages of general greeting situations. Their research shows that body movements, gestures, facial expressions, and salutations are useful signals for predicting human engagement. Concretely, (i) Sighting, orientation, and initiation approach; (ii) Movements and gestures that signal official acknowledges that a greeting sequence has been initiated. Smiles, waves, head nodes, *etc.* can all be part of this; (iii) Head dips; (iv) Approach, which assumes that the greeting process continues. In this stage, participants may gaze, groom, and extend one or both arms; (v)The final approach with smiling, mutual gaze; (vi) The close salutation at a standing position, including ritualistic speech and body contact. In addition, they also find that people's body behaviors change while approaching to initiate an interaction. Langton *et al.* also suggest that people use different cues to predict others' intentions based on distance, *i.e.*, using body posture, face orientation, and gaze direction from long to short distances [72].

Engagement during an interaction. Engagement is generally regarded as a state or a process. According to Oertel *et al.*, this notion is ambiguous across different domains [73]. While in terms of the state, participants are either engaged or not engaged, by process the concept emphasizes how interactors establish, maintain, and complete their perceived connection to each other during an interaction [74]. Note that the term *state* represents objectively observed facts in HHI or HRI, which is whether the participants are within interaction or not. It is used to distinguish itself from *process*. In this thesis, Poggi's definition of engagement

is adopted [75], which refers to the participant's inner state of being together with other participants and continuing the interaction.

Elements of Engagement. In general, engagement contains several elements [76–85]. They usually include behavioral, affective, visual, verbal, social, and cognitive signals.

- Behavioral engagement involves observable behaviors such as approaching, touching, avoiding, and hitting.
- Affective engagement is defined as the reactions that are usually represented by valence and arousal.
- Visual engagement encompasses actions involving the eyes and head such as maintaining contact or appearing inattention to others or materials.
- Verbal engagement reflects the sounds and semantic information towards other participants.
- Cognitive engagement refers to psychological investment and effort allocation of the person in order to fully comprehend the situation.
- Social engagement includes the activities of encouraging or disrupting others.

Moreover, these elements are not mutually exclusive but often overlap with each other.

Application of Engagement in social HRI. EE is studied in many disciplines and interaction scenarios. A simple taxonomy is based on the type of interactors: engagement in HHI or HRI. Although participants may behave differently in HHI and HRI, the estimation of engagement in these two disciplines is similar in terms of methodology. Thus the knowledge from the HHI could be applied to HRI [86]. In addition, the application scenarios of EE could be different. Examples are everyday conversations, healthcare, and learning situations. In different scenarios, engagement may have different dominance of its elements. As a result, the corresponding estimation methods are different, and it is difficult to make a fair comparison of different methods. There is no universal approach.

Engagement in Geriatric Psychiatry. In geriatric psychiatry, the concept and measurement of engagement are well established. For example, Cohen-Mansfield *et al.* proposed an Observational Method of Engagement for PwD [87], which was one of the most well-known tools that many studies have used to measure engagement [88]. Following this concept, Jones *et al.* developed the Engagement of a Person with Dementia Scale (EPWDS) towards psychosocial activities by assessing the behavioral and emotional expressions and responses [89]. Perugia *et al.* presented an effective computing framework that specifies the components of engagement in HRI [85].

Moreover, robotic and computer assistance has been shown to be effective interventions. Moyle *et al.* designed a robot seal for PwD [5]. They found that the robot seal was more effective than usual care in improving mood states and agitation and participants were more engaged with it than with a toy. Similarly, Feng *et al.* introduced an interactive system involving a display and a robotic sheep to engage PwD [90]. They claimed that multi-modal stimuli played a significant role in promoting engagement.

2.2.3 Personality Theories and Applications

Personality is used to describe a person's character that reflects a set of behavior, emotion, and cognition patterns. To quantify personality, traits are used to provide multidimensional aspects of personality. Several trait theories exist, such as Big-Five Inventory (BFI) [91] and HEXACO [92], *etc.* In this work, either BFI or HEXACO are used according to the datasets. The related adjectives for BFI and HEXACO traits are listed in Table 2.1 [92, 93].

Big-Five Inventory includes five traits, openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). Specifically, the appreciation for unexpected ideas, inventiveness, and variety of experiences is referred to as openness. Conscientiousness is a tendency to exercise self-control, act responsibly, and strive for success. Extraversion is defined by a wide range of activities, a high level of surgency from external activity or situations, and the ability to generate energy through external means. Individual differences

in social harmony are reflected in agreeableness. The predisposition to experience negative emotions such as depression is referred to as neuroticism. Moreover, there are various versions of BFI such as NEO-PI-R [94], BFI-44 [95], and BFI-10 [96], differing in the design of short-phrase items and the number of items. Taking BFI-10 as an example, there are 10 phrases, two of which correspond to a trait from OCEAN, and each phrase will be marked with an integer score between 1 and 5 (or sometimes 1 to 10). The score for each trait is the sum of the two phrase scores. The final personality is represented as the combination of the five-dimensional traits.

HEXACO conceptualizes personality in six dimensions. There are overlaps between the HEXACO theory and the BFI, but there are also significant differences. Specifically, those who rank high on the honesty-humility scale have less desire for wealth and do not take advantage of others. High emotionality scorers are more likely to be fearful of danger, anxious about stresses, and in need of others' emotional support, but they are also more likely to build empathy and emotional attachments to others. High extraversion individuals are more assured, like social situations and relationships, and exhibit positive enthusiasm and energy. High agreeableness people are tolerant of suffering, forgiving of others' mistakes, and cooperative with others. High conscientiousness individuals manage their time well, keep their surroundings neat and orderly, pursue their goals with discipline, and aim for precision and excellence in their job. High scorers on the open-experience scale tend to be more creative, easily become addicted to the beauty of nature and the arts, have unrestricted imaginations, and are bursting with curiosity about a wide variety of novel experiences.

In addition, personality measures include both self-reported and peer-rated methods, which also correspond to real personality and apparent personality, respectively. Compared to real personality, apparent personality is considered less consistent and vulnerable to the environment, interlocutors, and subjective attitudes of raters. In computer science, researchers use the personality scores from the self-reported or peer-rated questionnaire as the ground truth by default, but it is important to note that these two ground truths do not necessarily match and have intra-agreement or inter-agreement.

TABLE 2.1: BFI and HEXACO personality traits with corresponding descriptions.

	Traits	Descriptions (adjectives)
BFI	O	curious, imaginative, artistic, wide interests, excitable
	C	efficient, organized, not careless, not lazy, not impulsive
	E	sociable, forceful, energetic, adventurous, enthusiastic, outgoing
	A	forgiving, warm, not stubborn, not show-off, sympathetic
	N	tense, irritable, not contented, shy, moody, not self-confident
HEXACO	H	sincere, honest, faithful, loyal, modest/unassuming
	E	emotional, oversensitive, sentimental, fearful, anxious, vulnerable
	X	outgoing, lively, extraverted, sociable, talkative, cheerful, active
	A	patient, tolerant, peaceful, mild, agreeable, lenient, gentle
	C	organized, disciplined, diligent, careful, thorough, precise
	O	intellectual, creative, unconventional, innovative, ironic

2.2.4 Emotion Theories and Applications

Emotion is a collection of psychological states including personal experience, behavior, and peripheral physiological responses, which is one of the core features of the human mind [97]. In psychological science, researchers have proposed many different theories of emotion models, which could be categorized into three types, (i) category theory, (ii) dimension theory, and (iii) wheel theory. A classical model in each of the three types of theories are shown in Figure 2.4.

Category theory defines emotions into discrete categories. One of the widely used category theories is Ekman's six basic emotions, or sometimes seven with a natural state [98]. The six basic emotions consist of anger, disgust, fear, happiness, sadness, and surprise, which are the most common, culture-independent facial expressions. However, this assumption does not always work, *i.e.*, basic emotions can be interpreted differently depending on cultural background. In addition, although this is the easy-to-practice theory in computer science, categorization becomes complicated when expressing nuanced emotions.

Dimension theory is proposed to address the issues with category theories. Basically, the concept is to project emotions in a continuous space, such as valence-arousal (VA) space [8] or pleasure-arousal-dominance (PAD) space [99]. In the VA model, valence gives a positive or negative direction to the emotional state,

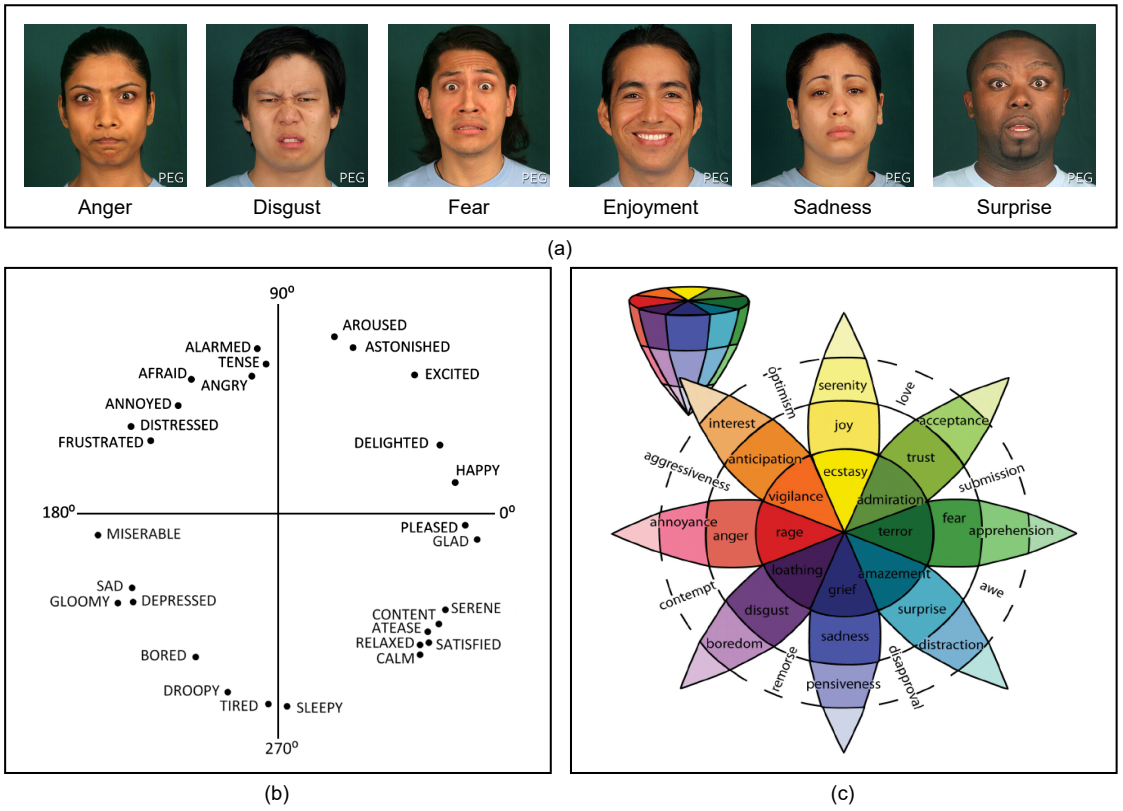

FIGURE 2.4: Three types of emotion theories: (a) category theory, taking Ekman's basic emotion as example [7], (b) dimension theory (Russel's model of affect) [8], (c) Plutchik's wheel theory [9].

and arousal measures the intensity of the feeling. PAD adds another dimension called dominance, measuring the control level of the situation, or in other words, expressing how one influences the surroundings and other people, as well as how his/her is influenced by them.

Wheel theory can be considered as a combination of category and dimension theories. Eight fundamental emotions and the level of emotional intensity are symbolized by a cone [9]. The intensity of emotion is described by the vertical axis of the cone, that is, the intensity of emotion decreases as it moves from inside to outside of the wheel, as shown in Figure 2.4.

2.3 Engagement Intention Estimation in HRI

In the area of HRI, the concept of engagement was investigated from many aspects [84], including the engagement intention estimation before a real interaction [34, 36–41], robot behaviors for attracting people to engage [100], the analysis of the engagement social cues and human affect [101–103], the early detection of engagement breakdown [104], the engagement estimation during interaction [1, 105, 106], and engagement density controlling in multi-party conversation [107]. HRI is a continuing process, along with user engagement. For the target problem, user engagement intention estimation happens prior to the start of the interaction. Figure 2.5 presents a human-robot interaction process and corresponding engagement research problems.

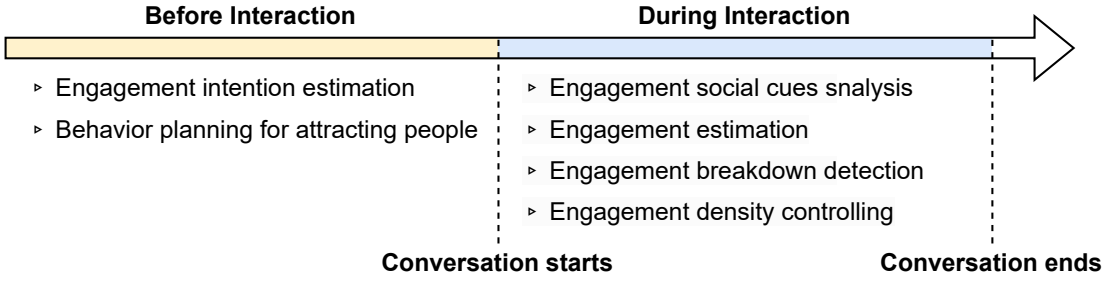

FIGURE 2.5: Human-robot interaction process. Research problems that are relative to engagement are categorized based on the time period.

In terms of user engagement intention estimation, Michalowski *et al.* propose a spatial model of engagement for a robotic receptionist situated in a booth near the entrance of an office [37]. The robot system gained spatial information from a laser tracker and the user's head pose information from a camera as input features, and then categorize users into one of the following four engagement groups: present, attending, engaged, and interacting. To some extent, their approach, as the pioneer of engagement intention estimation, is a relatively heuristic way that only considers the spatial features and the classification is based on observational behavioral analysis.

Bohus and Horvitz propose a machine learning approach to detect user engagement intention [36, 38]. The prediction model is embedded in an interactive kiosk that displays a realistically rendered avatar head. When a person is detected, (a) the

location of the detected face in the visual scene, (b) the width and height of the face region, which indirectly reflects the proximity of the agent, and (c) a probability score and a binary version which reflects the confidence that the face is frontal and thus provides an automatic measure of the focus-of-attention of the agent, are generated and forwarded into a pre-trained maximum entropy model to make a prediction. They also compute the trajectory of the above features and compare the prediction results of different feature combinations. Their results show that the prediction can be made 3-4 seconds before actual engagement with a false positive rate of 7.2% as a minimum.

Similarly, Xu *et al*. developed an engagement-aware robot using a machine learning approach [34]. They detected both engagement and disengagement intentions of users with a Support Vector Machine (SVM) using features such as the direction of attention, change of speaking status, change of emotions, and distance. They also detected the attention saliency of the users to estimate who is engaged and which ones have higher attention saliency and deserve more attention. To achieve this, they used a linear regression model considering attention saliency as a linear variable.

Vaufreydaz *et al*. propose an approach to detect engagement towards a companion robot using 32 multi-modal features including spacial, acoustic, and body features [39]. The adopted classification techniques are the Multi-Class Support Vector Machine (MC-SVM) and an Artificial Neural Network (ANN). They compare the performance of these two classifiers and find that detecting performance does not lose by reducing the features space to seven features: face size, the lateral position of the face in the visual scene, activated audio beam, audio localization, speed in the x-axis, position on the y-axis, and the relative orientation of the shoulder in the body. However, they do not explain the results and the reason for the contribution of these seven features.

Foster *et al*. compare a rule-based hand-coded method and supervised learning methods to classify user engagement in bartender-supporting interactions [40]. The rule-based method assumes that a user engages with the system when he/she comes close to the bar and makes eye contact with the bartender, while the data-driven approach uses the face and head coordinates of the user as input features. The

output class is either *notSeekingEngagement* or *seekingEngagement*. They point out that the performance of the rule-based method is even more competitive with that of the trained classifiers. Besides, they present a classifier based on conditional random fields (CRFs) to address the frequently changing prediction results.

Ozaki *et al*. adopt a deterministic finite automaton (DFA) method to solve the same problem [41]. The experiment is conducted in a reception system located at the entrance to a virtual company, with an RGB-D sensor and an IR distance sensor device. The real-world position and the head angle of the pedestrian are collected to encode the pedestrian's action. Particularly, the action is a set of binary values of gaze, near, far, exist, stop, approach, and leave. They also design seven pedestrian states from an observational study, including not found, passing by, look at, hesitating, approaching, established, and leaving. Finally, a state transition is made based on the original state and a rule-based transition function.

Existing weaknesses. Although most works have their contributions to managing engagement intention, most of the previous work is still heuristic. Specifically, different feature sets are used to reveal human engagement intention without general support from psychological and cognitive science [36, 37, 40, 41], after which collected data is fed into some black-box classifiers to make prediction [39, 40]. The applied machine learning classifiers are trained with the default parameters of corresponding toolkits, so the evaluation and comparison are not reliable enough. Moreover, temporal information from data is not adequately utilized. Some work predicts user engagement intention only uses the data from a single frame [37, 39] or simply calculated trajectory [36] to add time-related factors. Nevertheless, the temporal signal contains rich information that is fairly useful to make stable engagement tracking. [36, 37, 39–41] use features that are easy to gain such as body orientation to instead crucial cues like gaze direction, which has been proved as the crucial signals for engagement from psychological and cognitive science. Finally, The discussion of the multi-party interaction scenario is overlooked. Although [36] and [40] mention that their experiment scenario is multi-party interaction, none of them actually discuss the situation of multi-party. Their approaches are only be applied in a two-party conversation, *i.e*., there is no interaction between human participants. The estimation is made from the view of a robot system but ignores that new participants may express engagement intention towards another member

in an established interaction, which also can be inferred as having engagement intention with the robot.

The techniques employed by existing works can be grouped as rule-based and data-driven methods. With the progress in computing power and machine learning algorithms, classifiers such as SVM and VNNs are often used. However, data-driven approaches are still rare due to the lack of adequate datasets. Recently deep neural network approaches are proposed [108]. In this work, a multi-branch model is proposed, which learns from social signals and image features. While the deep CNNs [109] are used to extract image features, relatively shallow multi-layer perceptrons are used to learn the pattern in social signals that have a small dimension. Moreover, thanks to the LSTM [110], temporal information is exploited to improve the estimation stability.

2.4 Engagement Estimation in HRI

Traditional EE [111–115] extracts high-level social features, for example, body pose, facial expressions, and gaze, followed by a machine learning classifier. Recently, with the great progress of machine learning in computer vision, more and more deep learning methods have been developed [77, 78, 116–120]. A summary of the estimation methods is given in Table 2.2.

Machine Learning Classifiers. In general HRI, Salam *et al.* classified engagement using support vector machine (SVM) and random forest (RF), depending on predicted personality in a triadic interaction [111]. They advanced the concept of engagement to the group level and claimed that categorization of engagement based on individual and interpersonal features without personality is insufficient. A similar work was proposed by Celiktutan *et al.* [112]. Ben-Youssef *et al.* studied engagement from the breakdown perspective, *i.e.*, users leave before the expected end [113]. They extracted non-verbal multi-modal data such as the distance to a robot, gaze and head motion, facial expressions, and audio. A logistic regression (LR) classifier was used.

Another widely investigated situation is online or in-class learning. Monkaresi *et al.* explored engagement in the situation where students completed an online writing activity [114]. Heart rate, action units (AUs), and local binary patterns were extracted and fed to some classifiers like Naive Bayes (NB). Gao *et al.* predicted students' learning engagement in real-world classes on emotional, behavioral, and cognitive levels [115]. They used a set of features from wearable and indoor sensors to infer students' engagement status.

Deep Neural Networks. The aforementioned approaches require the expert design of input features and cannot efficiently deal with large feature dimensions. Del Duchetto *et al.* proposed a regression model based on CNNs and Long Short-Term Memory (LSTM) networks, which allows robots to compute the engagement from ego-view HRI videos [117]. The model was built on a long-term dataset from an autonomous tour guide robot in a museum. Zhu *et al.* presented an attention-based Gated Recurrent Unit network to predict engagement of students learning online [118]. Taking the advantage of the published dataset [113], Saleh *et al.* applied Inflated 3D ConvNets architecture to predict engagement state in an end-to-end way [116].

To estimate the engagement of children with autism spectrum disorder interacting with robots, Anagnostopoulou *et al.* compared AlexNet [121] and 2D CNNs using 2D or 3D poses [120]. Rudovic *et al.* proposed personalized reinforcement learning (RL) to estimate engagement levels (low, medium, high) from videos of child-robot interactions [119]. The videos were labeled offline by experts, and used to personalize the policy and engagement classifier to a target child over time.

For HHI, Sumer *et al.* utilized video recordings of classes to get attentional and emotional engagement features, and then applied SVM, RF, multilayer perceptron (MLP), and LSTM to predict students' engagement levels [78]. Guhan *et al.* described a multi-modal GAN-based approach, called ABC-Net, to identify engagement from online dyadic HHI recordings [77]. They utilized three-branch networks to gain valence and arousal, from which they generated engagement labels.

Existing weaknesses. All the previously mentioned methods require expert involvement. To achieve automated estimation, Steinert *et al.* proposed a vanilla LSTM model to predict emotional engagement [122], based on facial features (by

TABLE 2.2: Comparison of Engagement Estimation Methods.

Paper	Scenario	Participants[1]	Modality(s)[2]	Approach [3]	Output[4]
[113]	HRI	I/G	vis, aud	LR	$\hat{y} \in$ {NBrk, Brk}
[115]	HHI	G (age 15-17)	phy, env	LightGBM	$\hat{y} \in [1, 5]$
[114]	HCI	I (age 20-60)	vis	NB	$\hat{y} \in$ {Eng, NEng}
[111]	HRI	M	vis, dpt, per	SVM & RF	$\hat{y} \in$ {Eng, NEng}
[120]	HRI	I/G (children)	vis, dpt	AlexNet & 2D CNNs	$\hat{y} \in$ {Eng, MEng, NEng}
[117]	HRI	I/G	vis	CNNs+LSTM	$\hat{y} \in [0, 1]$
[77]	HHI/HCI	I	vis, aud, txt	GANs	$\hat{y} \in$ {Eng, NEng}
[119]	HRI	I (age 4-6)	vis	RL	$\hat{y} \in$ {HEng, MEng, LEng}
[116]	HRI	I/G	vis	I3D	$\hat{y} \in$ {Eng, NEng}
[122]	HCI	I (PwD)	vis	LSTM	$\hat{y} \in$ {Eng, MEng, NEng}
[78]	HHI	G (students)	vis	MLP & LSTM	$\hat{y} \in$ {HEng, MEng, LEng}
[118]	HCI	I (age 19-27)	vis	GRU	$\hat{y} \in$ {HEng, Eng, BEng, NEng}

[1] I, G, and M denote individual, group, and multi-party. The difference between multi-party and group is that multi-party treats participants separately but group treats them as a whole.

[2] Modalities: vis = visual, dpt = depth, per = personality, aud = audio, phy = physiological, env = environmental, and txt = text.

[3] Symbol '&' indicates using both and comparing with each other, and symbol '+' means combining to form a framework.

[4] $\hat{y}$ represents the inferred engagement label or value. For classification, Eng = Engage, Brk = Breakdown. The letters before Eng and Brk are N = Not, H = Highly, B = Barely, and M = Medium.

OpenFace [123] and VGGFace [124]) and contextual information (daytime, wellbeing, *etc.*) Similar to [122], this work is also an automated method. Not only visual facial features, but also affective features, behavioral features, and the relationships among all participants in the main interaction group are used in the research.

2.5 Personality Estimation in HRI

Many unimodal or multi-modal personality prediction approaches have been proposed with good estimation performance. Vision-based personality prediction aims to predict a target person's personality from pure visual modality, *e.g.*, RGB images or videos. The corresponding approaches are based on the assumption that personality traits lead to stable patterns of behavior that people tend to exhibit. Commonly used visual features are facial and body information. Through the analysis of these non-verbal behaviors, models are expected to provide scores of personality traits.

Preliminary studies have to extract well-designed features first, which requires professional knowledge from experts. Salam *et al.* introduce a personality regression method that utilizes both individual and interpersonal features [111]. They extract motion, skeleton activity, histogram of gradient, *etc.* as individual features and visual focus attention, relative orientation, distance, *etc.* as interpersonal features. Finally, a Gaussian process regression model is used to provide the final personality score.

Recently, more works rely on deep learning techniques. In ECCV ChaLearn LAP 2016 challenge [125], a competition to predict the first impression, *i.e.*, apparent Big Five personality traits from multi-modal YouTube videos, many competitive deep learning models are presented with the best accuracy above 0.9. Gurpinar *et al.* proposed an approach extracting face and scene representations from pretrained VGG-19 and audio representations from OpenSmile are fed into an extreme learning machine [126]. In contrast to using body language, Tellamekala *et al.* propose a facial-based regression approach. They propose an uncertainty-aware model that first identifies emotions and then uses the predicted emotion distributions and image features for personality recognition [127].

In the interaction scenario, Celiktutan *et al.* predict real and apparent personality by incorporating various multi-modal features, *e.g.*, visual, audio, and physiological, and use a standard support vector machine (SVM) [112] for classification. They demonstrate that using features from different modalities improves overall performance, but the advantage depends on the interaction scenario and which personality trait is predicted. Romeo *et al.* present a benchmark of apparent personality from body language [128], which is built on [112]. They classify personality into low or high subsets. They evaluate 3DCNN [129], 3DResNet [130], VGG DAN+ [131], and CNN+LSTM [132] in terms of precision, recall, F_1, and AUC.

In terms of real personality prediction, Shao *et al.* model the cognitive processes to predict real personality [133]. They use the facial landmarks and audio from the speaker to model the listener's facial reactions and personality by using a CNN-graph method. Palmero *et al.* highlight the importance of context [134]. From their proposed action transformer-based model, context information, *e.g.*, local scene, extended scene (videos from interlocutor), and meta-data (age, gender, *etc.*) consistently improves the performance, but they also claimed that none of the proposed feature combinations achieve a substantial improvement over the model only consider face features. Salam *et al.* propose a personalized model using gender and age as the profiling criteria with Neural Architecture Search (NAS) [135]. Curto *et al.* present a Transformer architecture to model long-term individual and interpersonal features in dyadic interaction and predict real personality [136].

Existing weaknesses. These works provide valuable discussions on real personality prediction, and also explore the effects of different modalities and contextual information. However, they only focus on real personality, while the relationship between real and apparent personality, as well as their correlation with varying information, is not well explained.

2.6 Emotion Recognition in HRI

Emotion analysis has been investigated from various perspectives. The prevailing method of emotion recognition involves analyzing facial expressions. Eleftheriadis

et al. proposed an architecture that detects joint action units (AUs) and performs facial feature fusion in a shared low-level subspace, which is learned by Gaussian processes [137]. Fabian *et al*. introduced an algorithm to recognize AUs, AUs intensities, and 23 emotion classes [138]. Arriaga *et al*. proposed a CNN-based framework for real-time face detection, gender inference, and emotion recognition tasks [139]. Since the above arts usually require clear facial features, Li *et al*. introduces the attention mechanism to overcome the negative effects of occlusion [140].

Although the mainstream way of emotion recognition is based on facial expression analysis, the inference becomes very challenging when the facial information is unclear or difficult to obtain. Also, from a psychological aspect, facial expressions are not the only channel to convey affects. As a result, some works have investigated emotion recognition with multi-modal data. For example, Schindler *et al*. conducted research on a limited dataset of non-spontaneous postures recorded under controlled circumstances, using the body pose to identify the six fundamental emotions [141]. Nicolaou *et al*. used the position of the shoulders in addition to the features of the face to identify fundamental emotions [142].

In addition to the crucial roles of the face and body features for emotion recognition, context is also an important factor. Kosti *et al*. emphasize the importance of considering the context for recognizing people's emotions in images [143]. They proposed a CNN model jointly analyzing the target person and the entire background environment to recognize emotions. Their proposed architecture consists of two branches and a fusion module, which takes account of body and image features. Similarly, Lee *et al*. presented a context-aware emotion recognition network [144]. Unlike the former, Lee used a face encoding stream instead of the former's body feature extractor, and adaptively fuse the face and context features. More recently, Mittal borrowed Frege's context principle from psychology. They designed a three-interpretation approach to recognize emotion in wild scenarios, using faces and gaits, semantic context, and socio-dynamic interactions and proximity among agents [145].

Existing weaknesses. Many issues have still not been thoroughly studied and addressed. First, rather than emphasizing a person's actual emotions, the majority of the aforementioned context-aware research concentrates on identifying perceived

human emotions. In addition, emotion recognition of participants in human interactions, where a person is no longer an isolated individual with a fixed moment but rather a person who changes over time and is influenced by interactions and other interlocutors, differs significantly from single-person emotion recognition in general scenarios. The link between personality and emotion, or the degree to which personality affects the creation of feeling, is a topic that has not gotten much consideration. To solve these problems, a novel method of emotion estimation is proposed. The most relevant work is from Zhang *et al.* [146]. They presented a multi-task framework for emotion and apparent personality estimation, but their method estimates apparent personality from emotion, which is opposite to the proposed data flow, from personality to emotions.

2.7 Summary

In previous sections, multi-party social HRI and non-verbal information are reviewed, as well as four particular estimation tasks, such as engagement intention, engagement, personality, and emotion estimation. In this section, a summary of existing weaknesses for estimating non-verbal information in HRI and potential directions for further research will be provided.

Social psychology studies. As discussed in Section 2.3 and Section 2.4, researchers used to rely on their experience to select social signals as input, without sufficiently studying the relevant literature in behavioral and social psychology science. The process of feature selection also needs expert involvement. Moreover, while some studies have investigated the contribution of different social signals, they have not been adequately tested in a wide range of scenarios. Finally, current social signals, which are considered cues of certain intentions or affective states, are concluded from human-human interaction. However, as human beings increasingly interact with various social agents such as robots and virtual humans, there is a need for further study on how human beings interact with these agents.

Multi-party social HRI. It is very important that multi-party social HRI involves more complex interactions than dyadic ones, where there is more than one human participant involved. In such scenarios, social signals can be difficult to

detect and interpret. Thus, analyzing multi-party non-verbal social information is crucial for developing practical HRI agents. For example, the four target tasks presented in this thesis have primarily been studied in dyadic interactions. However, in real-world settings, these tasks are often present in multi-party interactions, and their study in such scenarios has not yet been discussed. Therefore, it is necessary to extend the study of these tasks to multi-party scenarios to develop more comprehensive and natural solutions.

Multi-scale and multi-modal learning. Firstly, there are relatively few methods that utilize both high-level social signals and low-level image features to exploit the joint strengths. However, incorporating multi-scale hierarchical features has shown good performance in various tasks such as in [147]. Therefore, effective fusion of multi-scale features is a promising direction for further research. Secondly, as mentioned earlier, multi-modal fusion methods have been suggested for non-verbal information estimation tasks. Combining different modalities that can complement each other often leads to better estimation performance. However, many existing multi-modal approaches have not been as successful as expected due to various challenges. This indicates that there is still room for developing more effective strategies for multi-modal fusion.

Dataset and benchmark. One of the weaknesses in research on estimating non-verbal information in HRI is the lack of datasets. Most existing datasets are limited to dyadic interactions and do not cover the complexity of multi-party social interactions. Moreover, there is a need for more diverse datasets that consider different cultural and gender backgrounds. In addition, existing datasets are often collected in controlled lab environments, which may not reflect the complexity of real-world interactions. To address these limitations, one potential direction is to collect more large-scale, diverse, and multi-modal datasets. Another direction is to explore the possibility of synthesizing datasets using generative models, which can provide a larger and more diverse training set without requiring expensive data collection efforts. Additionally, available benchmarks to evaluate model performance are needed. Currently, there is a lack of standard evaluation protocols and datasets, making it difficult to compare the performance of different methods and track progress.

Efficient algorithm development. With the advancement of computer vision technology, many new techniques can be helpful for non-verbal information estimation in HRI, but efficiently integrating them into real-world platforms remains a challenge. Some lightweight models, such as [148, 149], are designed for embedded vision applications. To achieve a balance between latency and accuracy, it is important to explore how to effectively apply these models to downstream tasks and to design novel models that are tailored to HRI scenarios.

The aforementioned potential research directions are not exhaustive, and there are numerous other promising directions. Due to constraints on time and resources, this thesis aims to provide some solutions to the first three directions: social psychology studies, multi-party social HRI, and multi-scale and multi-modal learning. The subsequent chapters elaborate more specific approaches taken to address these gaps. Nonetheless, there is still much room for further research in the field of non-verbal information estimation in multi-party social HRI, and it is hoped that this thesis can inspire future research in this area.

Chapter 3

Engagement Intention Estimation for Multi-Party Social HRI

3.1 Introduction

The problem addressed in this chapter is to estimate whether people around an IA have an engagement intention to start an interaction with it. Compared to many existing works that detect engagement during human-robot interaction (HRI), this chapter focuses on the period before the starting of interaction. There are many reasons for choosing this problem as the first work in multi-party social HRI. Firstly, engaging at the start of an HRI, the politeness is attractive to users. Secondly, engagement prediction enables the robots to actively invite and naturally interact with the users in dynamic multi-party scenarios. Traditionally, users utilize a pre-defined activation command to trigger a robot's interaction manager and then start a conversation. However, in daily life, robots are also needed to actively greet or attract someone's attention, which can be treated as a manifestation of social awareness. For example, greeting and attention attraction can be applied to receptionists and street sellers.

Estimating human engagement intention is a challenging task. First, human behavior and social cues that express the intention are mainly non-verbal information that is difficult to recognize. Second, the behavior in which people express their

intention is various and sustains short duration, which increases the difficulty in learning a general estimation model. Third, available datasets were collected in different scenarios by different devices, so the data types are often not the same, which makes the evaluation and comparison hard.

Michalowski *et al.* [37] proposed a model using distance and head pose to classify users into one of the four engagement groups: present, attending, engaged and interacting. Utilizing spatial information with additional moving trajectory, Bohus and Horvitz [38] developed a machine learning approach to detect engagement intention. Vaufreydaz *et al.* [39] proposed to detect engagement towards a companion robot using 32 multi-modal features including spacial, acoustic, and body features. They compared the performance of support vector machine (SVM) with vanilla neural networks (VNNs) and found that detecting performance does not decrease by reducing the features to 7 ones. Foster *et al.* [40] compared a rule-based hand-coded method and supervised learning methods to classify user engagement in bartender-supporting interactions. They pointed out that the performance of the rule-based method could even be more competitive than that of the trained classifiers. They also presented a classifier based on conditional random fields (CRFs) to address the frequent changing of the classification results. Ozaki *et al.* [41] encoded user positions and head angles into actions and then predicted user decisions through a deterministic finite automaton. Recently, Ito *et al.* [108] proposed an end-to-end model using RGB images to predict the start of user interaction, in which the learning between oracle faces and facial keypoints is highlighted for improving the performance.

Most of the existing works focus on extracting social signals heuristically and validating feature effectiveness *e.g.*, feature sets used to reveal human engagement intention do not have ground from psychological or cognitive science. They also analyze engagement intention for each frame without adequate exploration of the temporal information. The applied machine learning classifiers are trained with the default parameters of the toolkits. Temporal information from data, which contains rich information for stable estimation, is not adequately exploited. The multi-party interaction is often overlooked. Studies are given from the view of the robot system in the dyadic interaction, but by this way, a new participant

who expresses engagement intention towards another member in an established interaction might not be classified as a potential participant.

Inspired by the success of deep learning in computer vision and natural language processing, this chapter introduces a method for estimating engagement intention using multi-branch neural networks. The hypothesis is that the proposed approach can effectively estimate engagement intention in multi-party HRI by leveraging both social signals and visual cues. To validate this hypothesis, experiments are conducted on three publicly available datasets, and the proposed method is evaluated using several standard metrics such as accuracy, balanced accuracy, and F1-score. Although there is no available benchmark, comprehensive ablation studies are performed to determine the contributions of the individual components of the multi-branch neural network. The results of these experiments demonstrate the effectiveness of the proposed method and provide insights into the factors that contribute to accurate engagement intention estimation in multi-party HRI.

The proposed model is designed to take in social signals and image features as input leveraging multi-modal features to incorporate rich information from both sources. The output is a probability indicating a person's intention to engage with the IAs. Social signals can be obtained through sensors such as laser trackers or extracted from images using state-of-the-art algorithms, whereas image features are extracted using CNNs. The model is designed with a multi-branch architecture that processes different inputs independently. For example, only the social signal-related branch is activated if social signals are available but images are not available due to privacy concerns. It's worth noting that human engagement intention typically doesn't change rapidly, so the model incorporates LSTM layers to leverage temporal information and improve engagement intention estimation over time. The main contributions of this chapter are:

- A novel architecture is designed to estimate the engagement intention of potential participants in multi-party HRI scenarios. This new architecture is characterized by its multi-branch and adaptable input structures.
- The findings from psychological research are reviewed and summarized to identify high-level social signals that can be combined with image features to serve as inputs.

- A new neural network based on CNN and LSTM is proposed to estimate the engagement intention of potential participants.
- A novel feature transition method is designed to interpret multi-party social signals, which is essential to enable robots to perform well in such scenarios.
- The experimental results indicate that the proposed approach has good performance in terms of accurately estimating engagement intention by utilizing multi-modal features.

3.2 Overview of the System Architecture

3.2.1 Socially Intelligent Agent Platforms

In order to clearly explain the process of data collection, multi-party social HRI scenario setting, *etc.*, a brief explanation of the experiment platforms is presented first. I would like to express my sincere gratitude to my colleagues for their invaluable contributions. The social robot and virtual human they have developed has enabled me to carry out my experiments smoothly and efficiently.

In this chapter, a virtual agent, named Nicole, is used for engagement intention estimation. The appearance of the virtual agent is illustrated in Figure 3.1 (right). It should be noticed that a humanoid robot, called Nadine, is used in another task (Chapter 4), which is also shown in Figure 3.1 (left). Since both platforms have the same technical structure, for the sake of streamlining the presentation, the structure and functionality-related information of the social agents will be described here.

Nadine is a believable-looking female humanoid robot with synthetic skin. Natural facial emotions and body movements can be made by 27 degrees of freedom of her body and face. Nicole, a female virtual human, is created by the Institute for Media Innovation (IMI). Her appearance and animations are created in 3ds Max, and the behavior is generated in Unity.

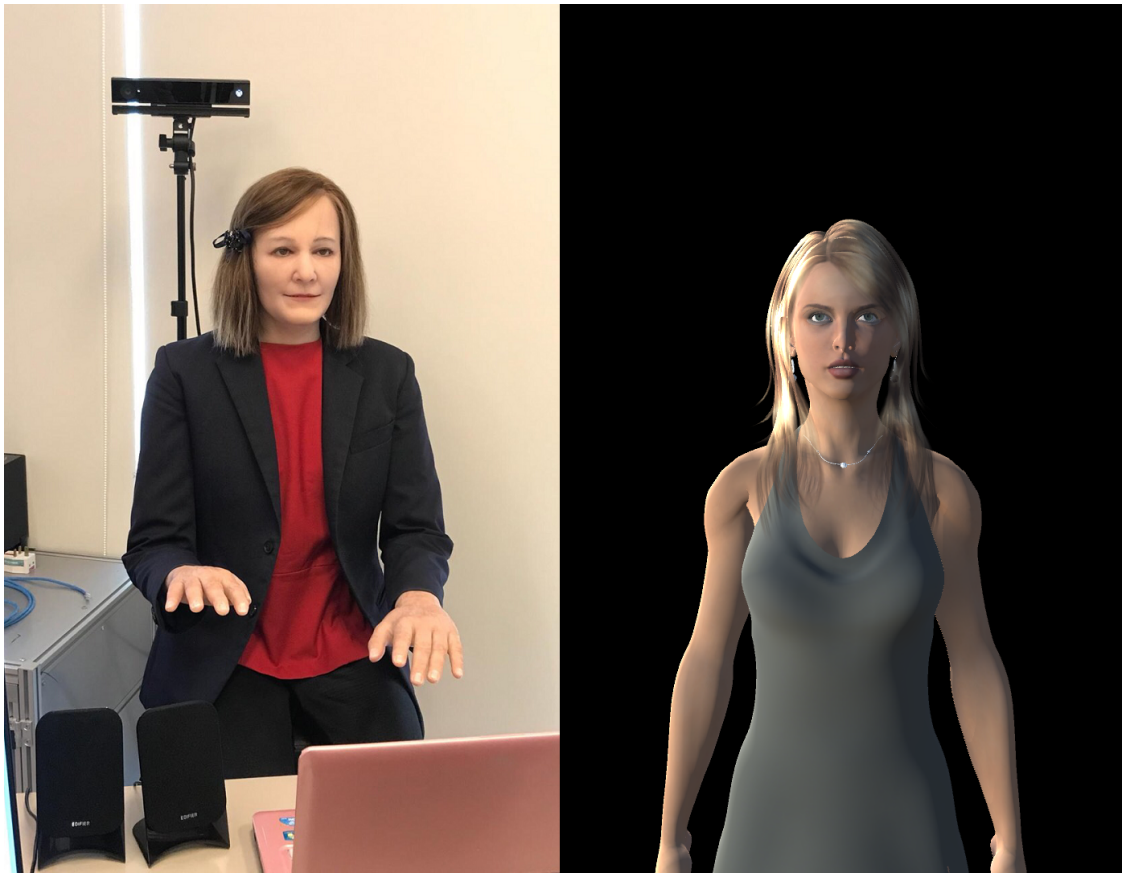

FIGURE 3.1: The appearance of Nadine robot (left) and Nicole virtual human (right).

The entire system includes three subsystems, namely (i) perception subsystems, (ii) processing subsystems, and (iii) action subsystems. Figure 3.2 shows the system architecture of the social companions.

The perception subsystem consisting of several sensors receives signals from the real world. Specifically, signals of scene image, depth, and participant skeletons are collected from a depth camera (Orbbec or Kinect V2), and the audio is gained through a microphone. After the collection of raw data, further processing to get semantic information is also done in the perception module, such as person identification, skeleton detection, and speech recognition. The semantic information, jointly with the raw data, is the input for the processing subsystems.

The processing component simulates the human cognitive and affective processes. From the peripheral area to the central area, managers and chatbot first focus on individual corresponding tasks, and then the processing results are forwarded to

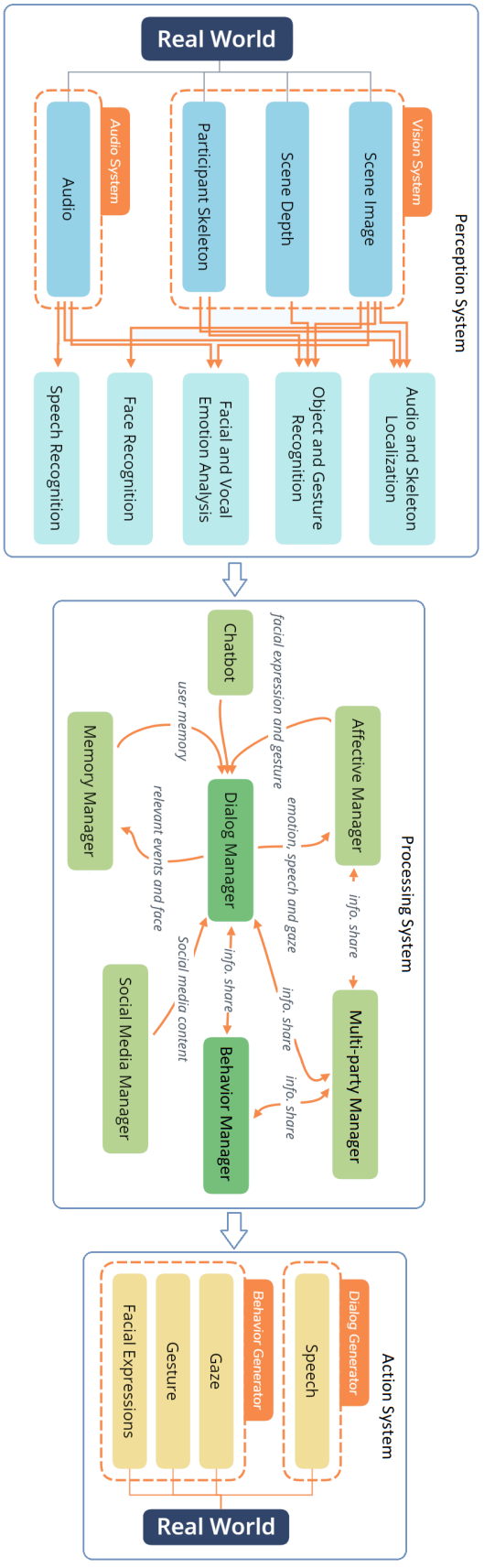

FIGURE 3.2: The system architecture of the virtual agent.

core managers, *i.e.*, the dialog and behavior managers, which are directly linked with the action subsystems.

Finally, the verbal language and non-verbal behavior are shown through the output hardware, such as synthesized speech and tone, head movement, body gestures, and facial expressions. An integrated interaction platform (I2P) is created to provide flexible integration and communication among different modules. The Thrift server architecture is used for data exchange and event posting across modules. C++, Java, and Python are utilized for system development.

3.2.2 Experiment Scenario

The experiment scenarios for studying engagement intention in HRI are various from different research groups. Some of them focus on a dialog system, a robot bartender, a robotic receptionist, or other functional virtual agents. In this work, the scenario is designed as a small group conversation including two human beings and a virtual agent. The three-participant scenario is a big change in HRI because these three parties form the minimum multi-party interaction. In the data collection and testing period, these three parties naturally form a small, free-topic, face-to-face conversation group, where the virtual agent serves as one of the group members.

3.3 Proposed Approach

The overview framework of the proposed user engagement intention estimation mechanism is illustrated in Figure 3.3. The original input X includes two pieces of sequential information of a user: (i) social signals and (ii) RGB images, collected from sensors like laser trackers, motion captures, and cameras. It is worth mentioning that these two types of data are not necessarily always available. An input buffer is designed to check the accessibility of each type of these data. Then high-level social signals and image features are extracted at the feature extraction stage (see Section 3.3.2). Feature transition is performed in multi-party scenarios (see Section 3.3.3). The processed features X' are fed into a multi-branch network

(see Section 3.3.4), which outputs a probability p of the user having engagement intention. When p is greater than a threshold δ, the IA greets him/her.

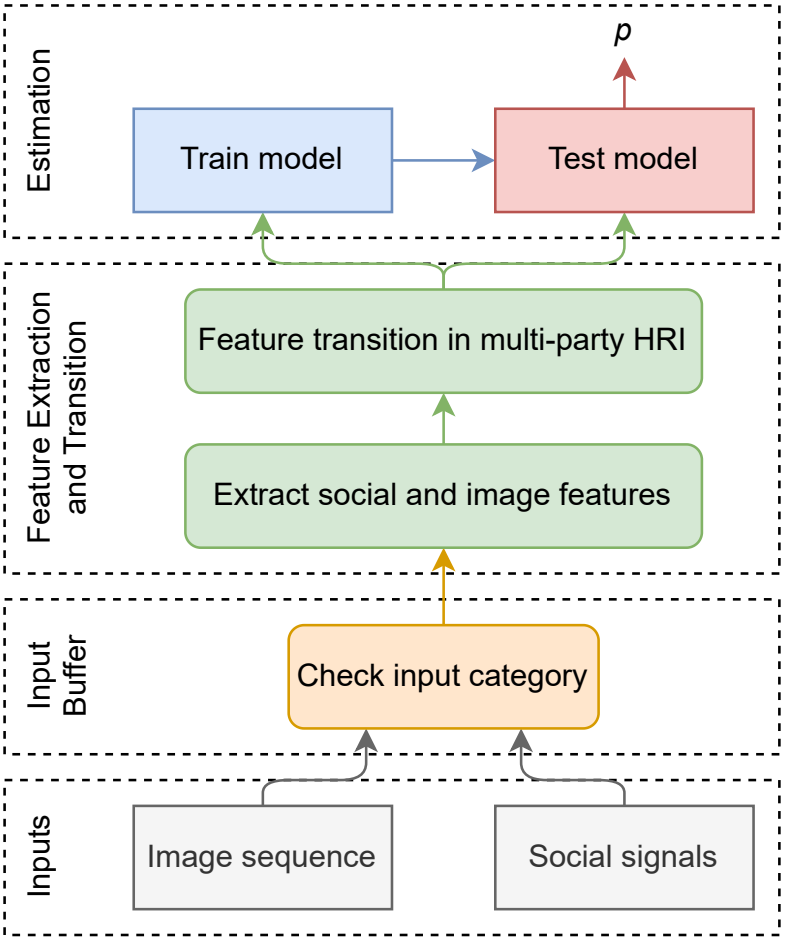

FIGURE 3.3: The framework of the proposed user engagement intention estimation.

3.3.1 Social Signal Selection

Proxemics is a common social cue in HRI [70] indicating the human use of space and the effects that population density has on communication and social interaction. Hall *et al.* introduced a spatial model that categorizes the surrounding area of a person based on different levels of social activities [69]. Studies [37, 38] proved the usefulness of distance and moving speed in engagement intention estimation.

Kendon and Ferber [71] suggested that there is a pre-phase when people want to initiate an interaction with others: sighting, orientation, and approach; official

movements and gestures; gaze, and extension of one or both arms; smiling and mutual gaze; ritualistic speech and body contact. Also, Langton *et al.* discovered that people gain social evidence from different cues based on distance changing [72]. Body orientation, face orientation, and gaze direction are used from far to near. Moreover, facial expressions is frequently used in multi-modal emotion and engagement studies [114].

Based on these studies, a set of engagement-related social signals are selected and categorized into (a) proxemic signals, (b) body signals, and (c) facial signals as shown in Table 3.1.

TABLE 3.1: Social signals for engagement intention estimation.

Category	Features
Proxemic signals	Location
	Distance
	Moving speed
	Moving orientation
Body signals	Body joint positions
	Body orientation
	Hand-waving state
Facial signals	Facial landmarks
	Facial action units (AUs)
	Face orientation
	Gaze direction

3.3.2 Feature Extraction

The raw input data may include social signals and images collected by sensors. Here, the extraction of social signals from images will be discussed. The subscript (p, t) denotes a person p at time t.

Proxemic signals. Proxemic signals of person p include his/her 3D location $l_{p,t}$, the distance $d_{p,t}$ between p and IA, his/her moving speed $ms_{p,t}$ and moving orientation $mo_{p,t}$. When these proxemic signals are not available, the 2D location of the body bounding box, the size of the bounding box, and its center moving speed are used to estimate the corresponding features.

Body signals. Body signals include body joint positions $j(n)_{p,t}$ where n is the joint index, body orientation $bo_{p,t}$, and hand-waving state $hw_{p,t}$. Similar to proxemic signals, the size of the bounding box, as a substitute, are employed to estimate body orientation. For body joints and hand-waving, the OpenPose network are applied, a toolkit for detecting the 2D pose of multiple people in an image [150]. A total of 25 body joints are extracted. Hand-waving state ($hw_{p,t}$) is generated from hand and elbow data, represented in the binary form as

$$hw_{p,t} = \begin{cases} 1, & \text{if } \ j(\text{hand}).y_{p,t} > j(\text{elbow}).y_{p,t} \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

where $j(\text{hand}).y$ and $j(\text{elbow}).y$ represent the vertical locations of the hand and elbow.

Facial signals. Facial signals include facial landmarks $fl_{p,t}$ and AUs $au_{p,t}$, face orientation $fo_{p,t}$, and gaze direction $g_{p,t}$. Facial Action Units (AUs) are a standard way to objectively describe human facial expressions. OpenFace is an open-source toolkit [123] for analyzing facial landmarks, head pose, gaze, and AUs. As it has been successfully used for affective-related tasks, it is selected to extract facial signals. For facial landmarks, 68 2D key points are detected. To estimate facial expressions, 35 AU features are extracted. Face orientation is represented by pitch, yaw, and roll. For gaze direction, 3 directions for each eye in world coordinates are gained, along with a 2D averaged 2-eye gaze direction. To sum up, a total of 182 face signals for every frame are extracted.

Image features. To extract general image features, CNNs are utilized. Specifically, image features from the body and face images are separately extracted. To reduce the training cost, two pre-trained models are utilized to extract image features. OpenPose network is used to extract the body heatmap by removing the output layer. VGGFace [124], a deep CNN network that has been used for face recognition, is used for extracting face image features.

3.3.3 Feature Transition in multi-party social HRI

In the multi-party scenario, engagement intention estimation becomes more challenging. Firstly, data extraction and tracking are difficult because multi-person alignment is more complicated and skeletons might overlap due to occlusion or truncation by image boundaries. Secondly, the interaction turns into a complicated dynamic process where people can interact with each other rather than with the IA. Thus, in the multi-party scenario, the key is how to enable the IA to recognize a person's social signals that are expressed toward another participant.

Figure 3.4 shows two common multi-party social HRIs where sensors are placed above an IA in ego view. When the IA is talking to p_1, a potential participant p_2 is detected. In Figure 3.4 (a), the social signals from p_2 expressed towards the IA can be processed by the engagement intention estimation model in a robot-centered way. However, the robot-centered method cannot be directly applied to all multi-party situations. For example, in Figure 3.4 (b), the system setting is unchanged (*i.e.*, sensors are still placed above the IA) but p_2 indicates to p_1 that he wants to interact with her or join the conversation. Then the previously extracted social signals that are processed by the robot-centered way may not accurately convey p_2's real intentions. For instance, p_2 does not look at the IA, but gazes at p_1, which is a strong social signal of having engagement intention. However, the robot-centered data processing will interpret this signal as a big gaze deviation, which is a strong social signal of no engagement intention.

A feature transition is proposed to make the intention estimation be adapted to multi-party scenarios. As shown in Figure 3.4(c), the transition functions are controlled by participants' geometric relationships (location and orientation). Particularly, proxemic signals and joint positions are recalculated from the view of p_1. For example, distance $d_{p_2,t}$ is computed as

$$d_{p_2,t} = \text{sqrt}\left((l.x_{p_1,t} - l.x_{p_2,t})^2 + (l.z_{p_1,t} - l.z_{p_2,t})^2\right) \tag{3.2}$$

where $l.x$ and $l.z$ represent the x and z coordinates of the location.

Let o denote the orientation-related features, ia represent the IA, and θ be the angle formed by relative positions (Figure 3.4(c)). Then, the orientation of p_2 with

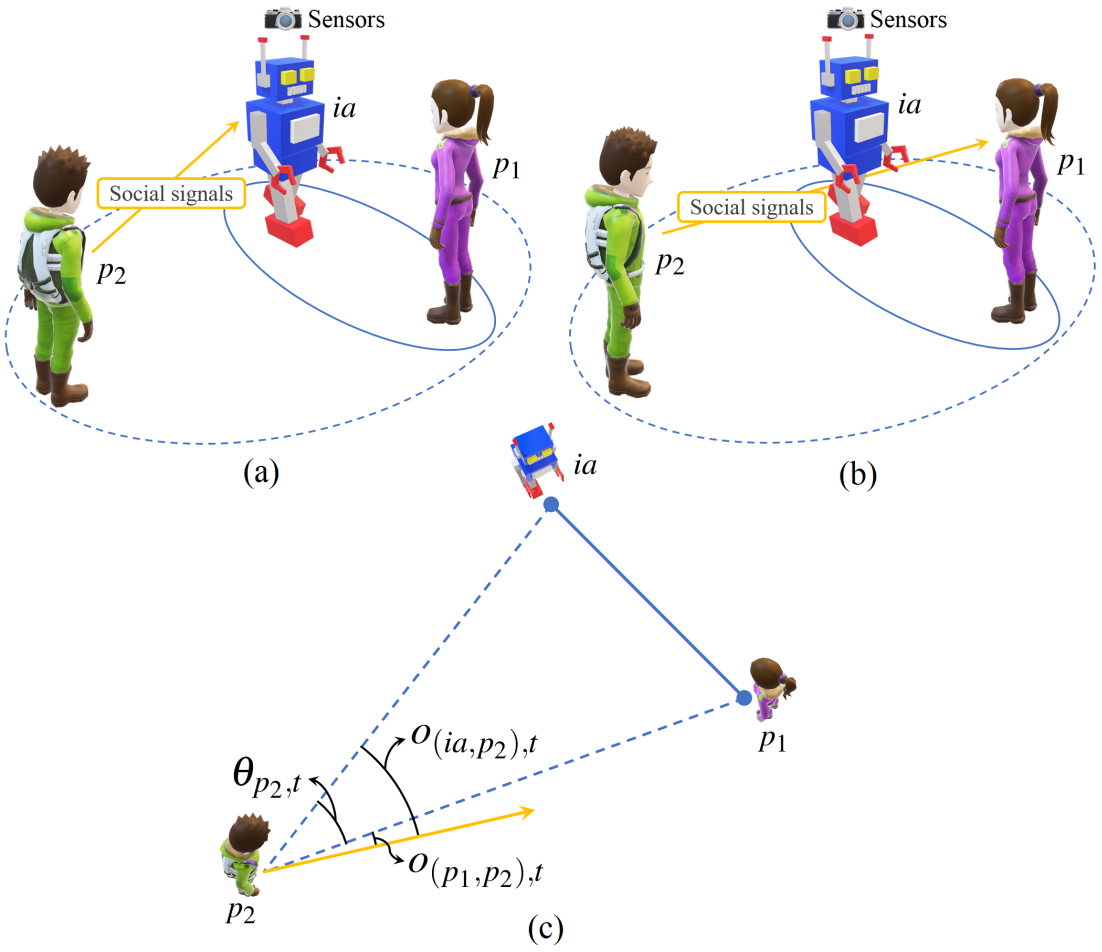

FIGURE 3.4: (a) and (b) are examples of multi-party social HRI. Dot and solid circles on the ground denote existing interaction and potential interaction, respectively. The yellow arrows denote social signals. (c) is the top view of (b) illustrating the transition of orientation-related features in multi-party social HRI.

respect to p_1 can be computed by

$$o_{(p_1,p_2),t} = o_{(ia,p_2),t} - \theta_{p_2,t}. \tag{3.3}$$

3.3.4 Engagement Intention Estimation

To estimate the engagement intention, a CNN-LSTM network is proposed, whose architecture is shown in Figure 3.5. The input X is a sequence of raw data x_t, $t = a, ..., a + T$ where a denotes the first frame index. After the data capturing process, three branches P, B, and F are designed to discover information

from three categories (proxemics, body, and face). Specifically, proxemic signals are sent to branch P. Body signals and body images are sent to branch B. Facial signals and face images are fed into branch F.

Five signal sets are separately forwarded into the CNN-LSTM networks. Only images are fed into the first CNN layer. P, B_s, B_i, F_s, and F_i denote the feature forward paths. The independent processing of each feature set has advantages. First, note that not all five feature sets are always available. Separating branches allows certain feature sets not to be required by easily deactivating the corresponding branches. Second, social signals, especially proxemic signals, innately have a smaller dimension size, but contain important semantic information. The multi-branch structure gives them a shallow network, which has the benefit of avoiding overfitting.

For branches B and F, the outputs from the middle fully connected layer (FC) are concatenated together followed by being forwarded to another FC layer. The objective is to learn the relationship between high-level social signals and general image features. The outputs from branches P, B, and F are denoted by $pp_{p,t}$, $pb_{p,t}$, and $pf_{p,t}$, which are the branch-level inferences of person p having engagement intention at time t. Finally, the probability $p_{p,t}$ is outputted by averaging three intermediate probabilities.

For each input $X = [x_a, ..., x_{a+T}]$, there is a corresponding label $Y = [y_a, ..., y_{a+T}]$. $T+1$ subsequent frames are used for training the LSTM to estimate the engagement intention of an individual target frame T, which means that the ground truth label is y_T. Considering that the datasets used in this chapter are highly imbalanced, weighted cross-entropy loss is adopted as the loss function L to train the model:

$$L = -\frac{1}{N}\sum_{N}\sum_{C} w_c y_{T,c} \log p_{p,T_c} \tag{3.4}$$

where N and C denote the mini-batch size and class, respectively, and w_c is the weight for class c.

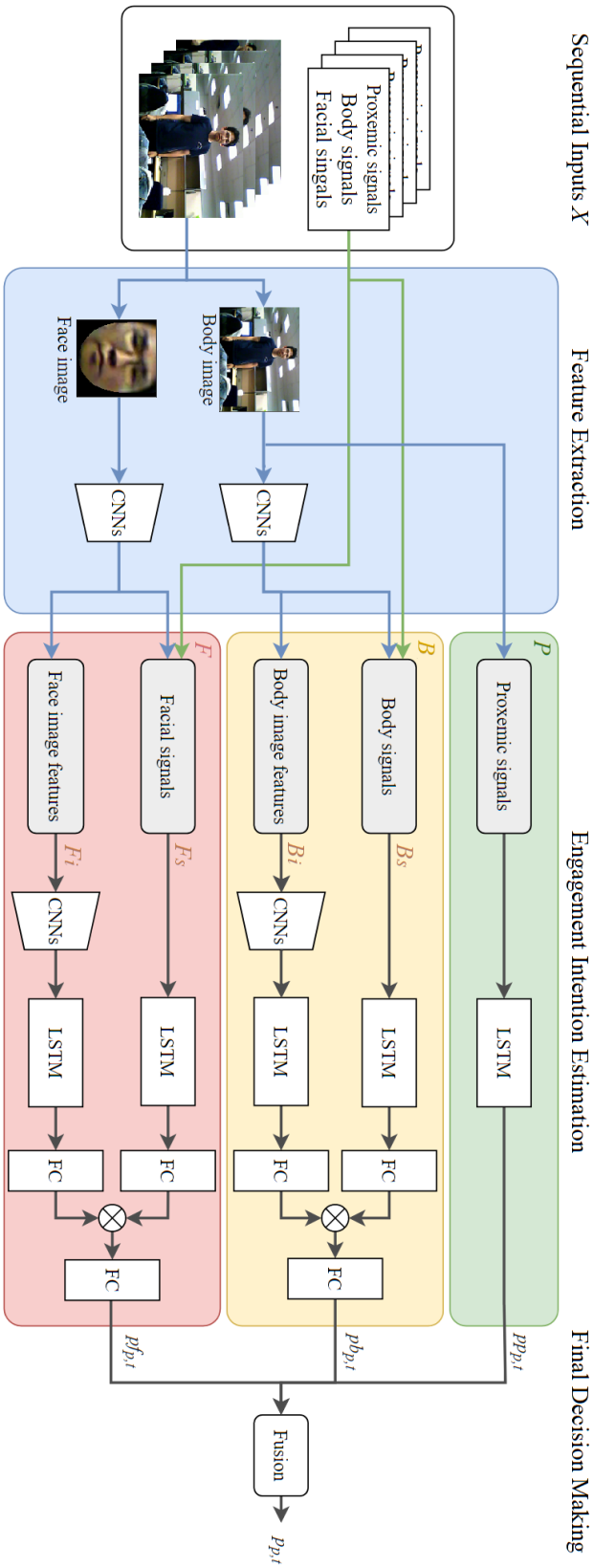

FIGURE 3.5: The architecture of the proposed CNN-LSTM network.

3.4 Experiments and Results

3.4.1 Datasets

Compared to estimating user engagement level during the interaction, the social signals used in the target problem are more imperceptible, momentary, and difficult to be captured, so the number and size of publicly available datasets are small. Also, existing datasets often include distinctive features, *e.g.*, raw images or high-level social signals. In order to evaluate the proposed approach, three public datasets are selected: ATC Trajectory [151], JPL-Interaction [152] and UE-HRI [153].

ATC Trajectory is a dataset for revealing the trajectory of people who want to talk to a robot, which was collected in the ATC shopping center in Osaka, Japan. The dataset includes 130 trajectories of people who intend to initiate interaction with the robot and people who do not. Features like position, moving speed, moving angle, and face angle are provided without videos or images. There is a total of 64 positive and 67 negative trajectories.

The JPL-Interaction dataset provides interaction-level human activity videos from the ego view. Shaking hands, hugging, patting, waving a hand, pointing, punching, and throwing objects are acted. Although JPL-Interaction is not designed for engagement estimation, it is a good dataset for target purpose since the videos last from the initial stage to the interaction starting. The dataset is re-annotated frame by frame and get 89 HRI sessions (46 sessions have engagement intention and 43 sessions have no engagement intention).

UE-HRI dataset was collected in spontaneous HRIs. It originally includes 54 sessions for studying user engagement breakdown, which has a wide range of data such as videos, depth images, tracked user position, head pose, and gaze, from heterogeneous sensors. Similar to JPL-Interaction, UR-HRI is not intrinsically designed for engagement intention estimation, but it contains data for a short period before people started to interact with the robot (47 positive samples) or when people just passed by (22 negative samples).

3.4.2 Implementation

Cross-validation are performed to test the proposed model. For each dataset, it has k interaction sessions. 80% interaction sessions are randomly chosen as the training set and the rest 20% are test set. The dataset is split on the interaction session level rather than the frame level. Each interaction session s has w_s frames $x_i,\ i \in 0, ..., w_s$ with corresponding annotation Y_k. The sequential input X from consecutive $T+1$ frames, *i.e.*, $X = [x_a, ..., x_{a+T}]$, are sampled. Then each sample X has an overlap of T frames with its next input.

The model is trained in PyTorch from scratch. The chunk size is set to $T = 29$ (the sequence length is 30 frames). In terms of the network architecture, a single-layer CNN is used with an output size of 32. The sizes of LSTM's hidden state for social signals and image features are set to 16 and 64 followed by a dropout layer with a rate of 0.5 to avoid over-fitting. The middle-stage FC layer's output size is 32 and the final-stage FC layer's output size is 2 with the Softmax layer. The batch size is set to 16. To optimize the cross-entropy loss, the model is trained for 20 epochs using the Adam optimization with an initial learning rate $\text{lr} = 0.0001$ and divided by 10 every 5 epochs.

3.4.3 Results

After the first-step exploration into the datasets, it found that the JPL-Interaction dataset is highly imbalanced, which is a very common issue in engagement intention estimation, so three metrics are adopted, namely accuracy, balanced accuracy (the sum of true positive and negative rates divided by two) and F1-score to evaluate the results. Table 3.2, Table 3.3, and Table 3.4 show the performance of the proposed approach.

In the ATC Trajectory dataset, the inputs include only social signals: proxemic and facial signals. Table 3.2 reports the performance of activating proxemic branch P, face branch F, and social signal paths P-F_s. It shows the benefit of using input features from different modalities. The method with the most input features achieves the highest performance on all metrics. Particularly, balanced accuracy

and F1-score are 0.788 and 0.808, respectively, which supports that social signals are assumed to represent people's engagement intention. Notably, the results from the proxemics branch are better than those from the face branch.

TABLE 3.2: Performance on ATC Trajectory dataset.

Method (Branches)	Accuracy	Balanced Accuracy	F1
Proxemic branch (P)	0.772	0.781	0.795
Face branch (F)	0.585	0.570	0.681
Social signals (P-F_s)	**0.779**	**0.788**	**0.808**

Table 3.3 shows the estimation results of the image-based dataset. After data extraction, all branches can be activated. First, the result from the body branch (F1=0.827) is better than that from the face branch (F1=0.465). This is contrary to psychology studies but reasonable in the computer vision area. During the implementation, it is found that the detection of facial features, especially gaze direction and eye landmarks, is challenging when the distance is large. When people turn faces or are far away from cameras, face detection may fail. To address this issue, it may be helpful to take into account the effect of distance and image resolution on the accuracy of feature extraction. Second, the estimation results of using social signals and image features are not close, which implies that both semantic information and image features are important. A similar pattern can be found from the results on UE-HRI in Table 3.4. The highest F1-score (0.873) is achieved by using all features, whereas the highest balanced accuracy is 0.801 when social signals are used. Moreover, it is found that using social signals tends to give more true negative samples but fewer true positive samples, which is contrary to general expectations. Therefore, in terms of better interacting with people who have engagement intention, using all features for engagement intention estimation seems to be a better choice.

3.5 Conclusion

In this chapter, a novel approach to estimate human engagement intention in multi-party social HRI is presented. The proposed approach leverages multi-modal features, including image features and social signals, to train a CNN-LSTM network.

Table 3.3: Performance on JPL-Interaction.

Method (Branches)	Accuracy	Balanced Accuracy	F1
Proxemic branch (P)	0.647	0.572	0.623
Body branch (B)	0.842	0.831	0.827
Face branch (F)	0.514	0.493	0.465
Social signals (P-B_s-F_s)	0.690	0.701	0.683
Image features (B_i-F_i)	0.746	0.761	0.739
All (P-B-F)	**0.851**	**0.844**	**0.887**

Table 3.4: Performance on UE-HRI.

Method (Branches)	Accuracy	Balanced Accuracy	F1
Proxemic branch (P)	0.581	0.505	0.712
Body branch (B)	0.649	0.675	0.754
Face branch (F)	0.605	0.561	0.727
Social signals (P-B_s-F_s)	0.704	**0.722**	0.798
Image features (B_i-F_i)	0.783	0.590	0.871
All (P-B-F)	**0.792**	0.662	**0.873**

The selection of social signals is based on the literature of social behavior science. The multi-branch architecture of the network allows it to adjust when different types of inputs are fed. Furthermore, the LSTM layer exploits temporal information to provide a stable estimation. Also, a novel feature transition method is designed to interpret multi-party social signals. Through a series of experiments and analyses on three datasets, the effectiveness of the proposed approach is demonstrated. The approach has the potential to equip IAs with the ability to proactively greet potential participants and enhance the user experience in multi-party social HRI scenarios.

Although the proposed approach can learn and estimate human engagement intention, it can be further improved. Face features like eye gaze and facial expressions carry critical information but are not fully delved into. People may have different behaviors if they take different roles like acquaintances or strangers, or are in different environments. Adding context to the estimation process will be helpful.

Chapter 4

Engagement Estimation During Multi-Party Social HRI

4.1 Introduction

This chapter considers the problem of the estimation of engagement of the elderly in wild multi-party HRI. With the advance of social robots, deploying robots at healthcare facilities becomes a possible solution to providing round-the-clock medical and psychological care to the elderly, especially the people with dementia (PwD), and supporting their caregivers as well [85, 154]. Natural elderly-robot interaction (ERI) helps make the robot a good companion for the elderly who usually experience declines in physical and cognitive capacities. This has a great impact since the proportion of people aged 60 years and older in the world will nearly double from 12% in 2015 to 22% in 2050 according to the World Health Organization [155].

During the ERI, if the robot can recognize the engagement state of the elderly, it helps the robot to respond to the elderly properly to maintain long-term interaction or to produce appropriate social behavior for the elderly to feel a sense of belonging. Here engagement refers to the inner state of a participant attributing to being together with the other participants and continuing the interaction [75]. Many studies have shown that engagement plays an important role in both human-human

interaction (HHI) and HRI [156]. There are many applications, such as therapy robots and virtual companions. This function can also be extended to more general scenarios, such as evaluating student engagement and online teaching.

Engagement estimation (EE) is a kind of affective computing and behavior recognition, and it goes further to probe the inner intention behind the apparent behavior and emotion. Many methods have been developed to estimate engagement in various scenarios such as general HRI [111–113, 116], museum tour guide [117], classroom or distance learning [78, 83, 114, 115, 118, 119, 157], and healthcare [4, 120, 122]. Conventional approaches use non-verbal cues such as proxemics, body pose, gaze patterns, facial expressions, and context information to build classifiers. Deep learning approaches have also been developed [77, 78, 116–118, 120]. However, most previous work assumes the interaction is in a laboratory environment or a dyadic situation. When the research is expanded to special populations and more complex circumstances as this chapter is (see Figure 4.1 for example), not much work has been done before. This may be in part due to the following challenges:

C1 The non-verbal signals from **the elderly** alter in facial shape and patterns of body behaviors along with aging [158, 159]. This challenges the conventional computer vision approaches in accurately estimating engagement state.

C2 From dyadic to **multi-party HRI**, understanding the dynamics and stability of the interaction becomes more complicated.

C3 In unconstrained **wild** space, moving people, bad lighting, confusing objects, *etc*. make it difficult to interpret the complex environment.

To tackle these challenges, a supervised learning method for EE from real-world multi-party ERI is proposed. The hypothesis is that by analyzing individual and group non-verbal behaviors, the proposed method can accurately estimate engagement in multi-party ERI. The effectiveness of the proposed method is validated through comparisons with existing approaches and the results of ablation experiments. Mean absolute error and mean squared error are employed to evaluate performance. Specifically, the method takes a video clip as input and outputs the estimated engagement state. Figure 4.2 shows the whole process. For each video clip, the behavioral, affective, and visual feature maps are extracted firstly. A

FIGURE 4.1: Three interaction sessions with five frames from the video recordings of real-world multi-party elderly-robot interaction demonstrate conversation dynamics (from one to more participants) and unconstrained environment (open space and free-moving background people). The videos are recorded from the robot ego-view, and $t = [100, ..., 4000]$ denotes five time stamps.

ResNet-3D [160] is adpoted as the backbone to generate behavioral features from the spacetime region. Affective and visual representations of participants are extracted using emotion recognition and face analysis tools. Then, these features are fed into an individual learning module, where features are refined by a self-attention mechanism. After that, the refined features are fed to the group learning module, which is a graph attention network learning the relationships among participants. The relationship conveys side participants' information, which helps the EE of the key elderly. Furthermore, to support the supervision, a real-world dataset is collected and labelled, which is the video recording of the interaction between the elderly and an intelligent social robot, using the conventional psychological approach. The main contributions of the chapter are

- A novel and automatic method for estimating participants' engagement levels in multi-party HRI is proposed.
- More challenging and less explored wild multi-party interactions between elderly people and a robot are analyzed, compared to existing approaches.

- A novel engagement estimation framework for such scenarios is designed by combining the engagement studies from psychiatry with computer vision techniques, where behavioral, affective, and visual engagement and their features are investigated.

- A new deep learning model is constructed, which includes self-attention networks and graph attention networks to learn individual and group information, which efficiently improves the performance of engagement estimation.

- A new dataset is created for studying engagement in multi-party elderly-robot interaction in the natural environment. The dataset includes multi-view video recordings and labeled annotations.

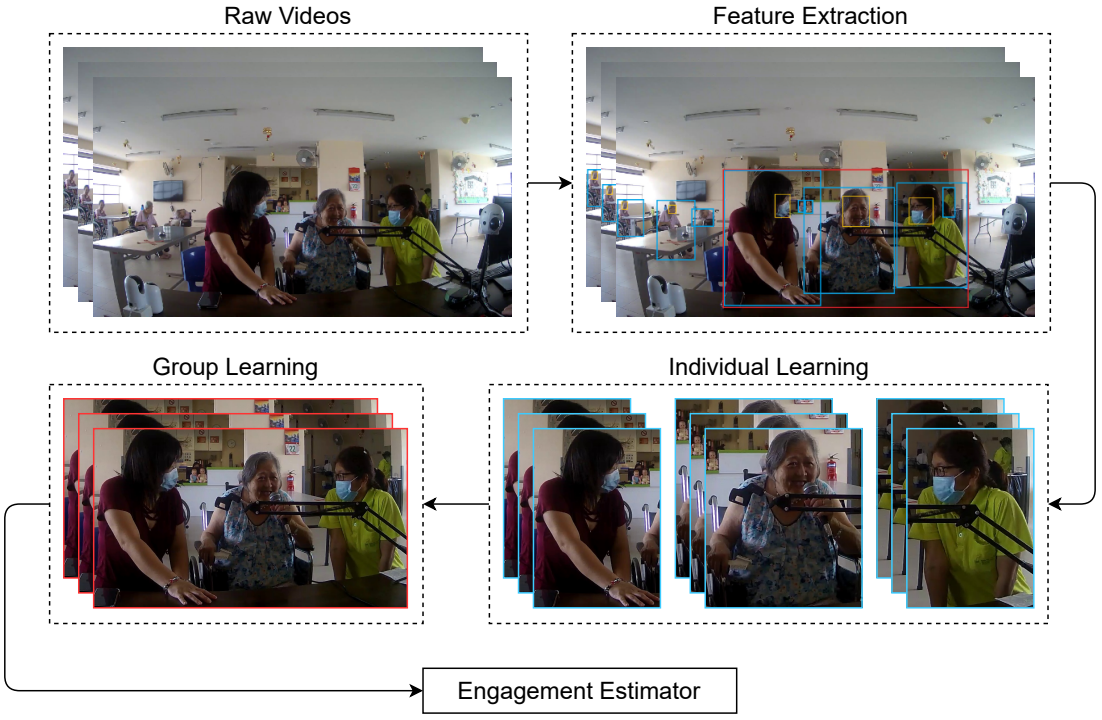

FIGURE 4.2: Overview of the proposed engagement estimation. The method is composed of four modules: (i) Feature Extraction, (ii) Individual Learning (Self-Attention Mechanism), (iii) Group Learning (Adapted Graph Attention Network), and (iv) Engagement Estimation.

4.2 Proposed Approach

According to Section 2.2.2, the elements of engagement include behavioral, affective, visual, verbal, cognitive, and social engagement. In this chapter, the goal is to estimate the engagement of the elderly interacting with a humanoid robot in casual conversation via a computer vision approach, so attention is paid to behavioral, affective, and visual engagement. The verbal element is eliminated due to the input modality, and the cognitive and social elements are overlooked due to the participant's physical and mental conditions.

The target problem in this chapter can be described as follows. In general, the goal is to estimate the value of engagement. The input is a raw video clip of the multi-party ERI in a wild environment. This video clip has a duration of about 10 seconds and contains only one elderly as the main participant and possibly a few other participants. It is assumed that within this duration the elderly's engagement state is fixed and corresponds to a value in $[0, 1]$. The value 0 represents the lowest level of engagement and the value 1 represents the highest level of engagement. In real applications, HRI usually contains several interactive sessions and each session can be further divided into a sequence of clips. If the value of engagement for each clip can be estimated, the engagement state over the entire interaction session can be obtained.

In this section, a supervised learning method for estimating the value of engagement of the elderly given a video clip is presented. The method consists of four modules. The first module is feature extraction. Some pre-trained network models is used to obtain spatio-temporal representations of the input videos, where behavior, affective and visual features are extracted. The second module is individual learning, which refines individual features by adding a self-attention mechanism. Because the facial and body features of older adults are difficult to recognize, a attention mechanism to enrich individual features is introduced. The third module is group learning. Here, a graph network is constructed to learn the response from nurses and the relationships of participants within the group, which further helps to understand the engagement of the elderly. The last module is engagement estimation, which generates a value representing the elderly's engagement state. This is done by a fully connected (FC) estimator, *i.e.*, the final layer of the network is a

FC layer that generates final output by regression. The loss function for training is based on the mean squared error (MSE):

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2, \tag{4.1}$$

where y_i is the predicted engagement value, $\hat{y}_i$ is the ground truth, and M is the number of video clips. Figure 4.3 gives the overall architecture of the proposed method. The first three modules are elaborated later in detail.

4.2.1 Feature Extraction

While there has been extensive research on human pose estimation and dynamic gesture classification [161, 162], a ResNet-3D [160] is selected as the backbone to capture spatio-temporal context of an input video clip. This is motivated by the promising performance of ResNet-3D models in a wide range of video-related benchmarks [163]. This backbone is pre-trained on Kinetics 400 [164]. The model details are listed in Table 4.1. The spatio-temporal feature maps, $X_{r4}^{B} \in \mathbb{R}^{1024\times4\times14\times14}$, extracted from res_4 layer represents the behavioral engagement. Here 1024 is the channel depth, 4 is the temporal dimension and 14×14 is the map size. As depicted in Figure 4.3, four feature maps are gained in time positions.

Meanwhile, multi-person tracking on each input video clip is performed. The purpose of this step is threefold: (i) the main interaction participants are identified based on the bounding boxes; (ii) these bounding boxes is used to eliminate the interference of redundant background information; and (iii) the tracked bounding boxes are used as constraints for face tracking as a way to ensure the consistency of face and body information.

Initially, multi-person tracking (MPT) is conducted using ByteTrack [165] to obtain the bounding boxes (BBX) of the detected bodies. The screening of the main interaction members is intuitively based on two variables: the frequency of a person's appearance in the temporal axis and the distance from the camera in the spatial axis. These two parameters can be viewed as an empirical design related to the experiment. The investigation results show that 20% and 5000 pixels are the

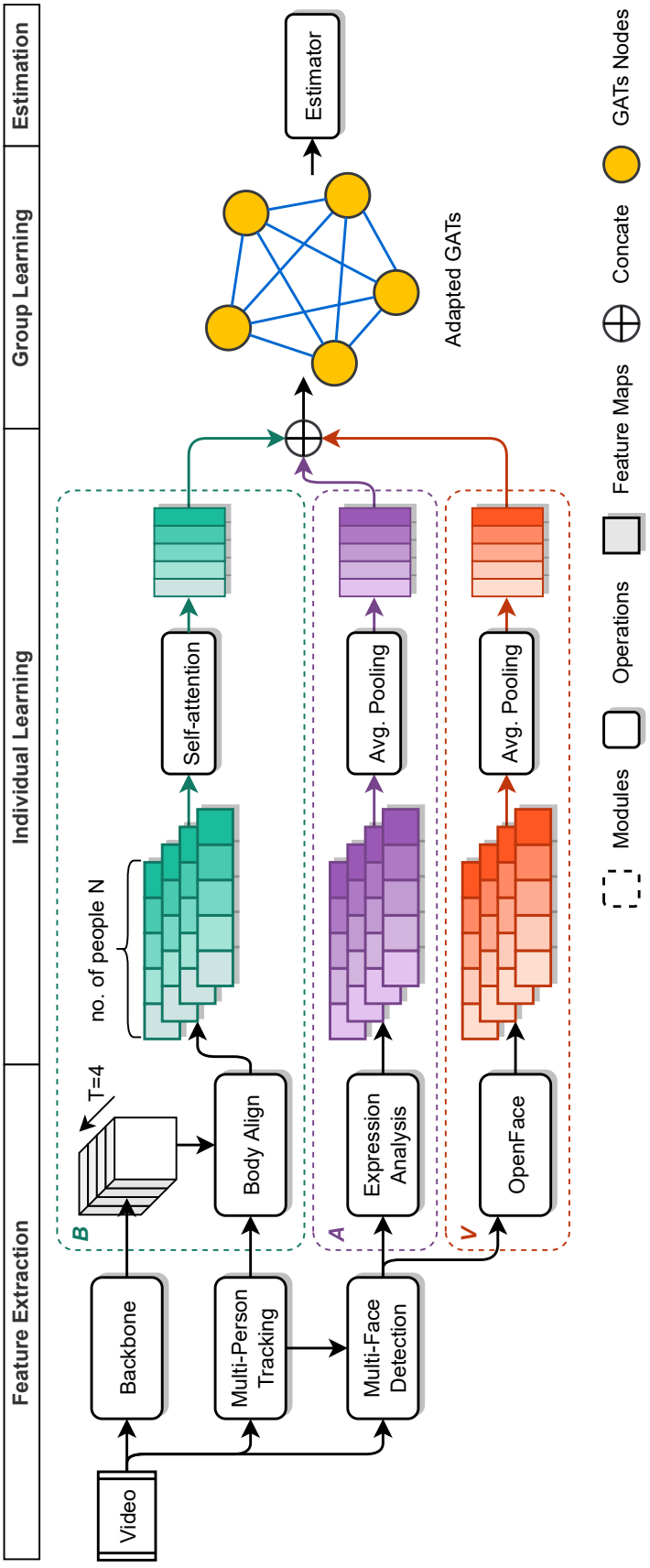

FIGURE 4.3: Architecture of the proposed network. A ResNet-3D model is adapted for extracting spatio-temporal features. Multi-person tracking and multi-face detection are used to get the bounding boxes in order to align and slice out corresponding body and face feature maps, which are then pooled for individual learning. The self-attention mechanism or average pooling is applied to refine the behavioral ($\mathcal{B}$), affective ($\mathcal{A}$), and visual ($\mathcal{V}$) features. The concatenation of these learned three components gives the representation of an individual, which is then further improved via group learning. Finally, a fully connected layer estimates the elderly's engagement state.

TABLE 4.1: The structures of the proposed backbone model and expression analysis module.

Layer Name	**Backbone**		**Expression Analysis**	
	Architecture	Output size	Architecture	Output size
conv1	$5 \times 7 \times 7$, stride 2, 2, 2	$16 \times 112 \times 112$	7×7, stride 2, 2	112×112
maxpool1	$2 \times 3 \times 3$, stride 2, 2, 2	$8 \times 56 \times 56$	2×2, stride 2, 2	56×56
res2	$\begin{bmatrix} 3 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	$8 \times 56 \times 56$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	56×56
maxpool2	$2 \times 1 \times 1$, stride 2, 1, 1	$4 \times 56 \times 56$	N.A.	
res3	$\begin{bmatrix} 3 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	$4 \times 28 \times 28$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	28×28
res4	$\begin{bmatrix} 3 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 6$	$4 \times 14 \times 14$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	14×14
res5	pruned		$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	7×7

optimal settings that are closest to real situation. The justifications behind this design are: (i) The interaction sessions in experimental data last about 3-38 mins. Lower frequency threshold (20%) is selected, because some side participants may play very important roles in a partial period of time; (ii) The distance between the camera and the participants is relatively fixed, so the size of the box can reflect the distance when the body shape of the participants is ignored. 5000 pixels is a choice that larger than most of the background characters and smaller than the main participants.

Given these, RoI Align [166] is used to project the coordinates on the frames' feature maps and slice out the corresponding features for each individual. After that, the behavioral feature maps are refined to $X^B \in \mathbb{R}^{N\times1024\times4\times7\times7}$, where N is the number of detected participants. Formally,

$$X^B = RoI\left(E\left(v\right), \mathrm{BBX}\right) \tag{4.2}$$

where v represents the video clip, and RoI and E are the RoI align and feature extraction operations.

Firstly, to extract affective and visual engagement features, multi-face detection is conducted by using RetinaFace [167]. In order to ensure the consistency of the face and body in the temporal dimension, the dimension of temporal direction is set to $T = 4$. The detected $N \times 4$ faces are used for affective and visual feature extraction.

Specifically, a pre-trained ResNet-18 model is employed as emotion recognition model, namely DMUE [168], on the cropped and aligned faces. To keep more facial information, the original 8 emotion classification results is not used, but instead, the final linear projection layer is removed and the mid-output from the res_5 layer is used to represent affective engagement. An average pooling is used to downsample features patch, which gives the affective features $X^A \in \mathbb{R}^{N\times4\times512}$.

Also, OpenFace [123] is utilized to extract the visual features. Since visual engagement is highly related to head and eye behaviors, head poses and gaze features are selected. Particularly, 6 head pose, 6 gaze directions and 2 gaze angles, 112

two-dimensional eye region landmarks, and 168 three-dimensional landmarks are extracted, which together form the visual engagement features $X^V \in \mathbb{R}^{N\times 4\times 294}$.

4.2.2 Individual Learning

For the behavior features extracted in the previous module, though they are localized to the bounding boxes, they lack detailed body posture and action information, which actually plays an important role in understanding behavioral engagement. To overcome this issue, a self-attention mechanism [169] is introduced to refine the behavioral features. The interaction between any two feature positions in spatio-temporal dimensions is expected to be learned by the attention mechanism, and this information is then used to enhance the feature representations by prioritizing the critical body regions in the spatial domain and frames in the temporal domain. As demonstrated by the ablation study in Section 4.3, capturing such fine details contributes to the improvement of estimation performance.

For implementation, the self-attention mechanism is a non-local operation, which calculates the response at a given position as a weighted sum of the features at all positions. That is, the self-attention block receives behavioral feature maps X^B extracted from the previous module as input and outputs the updated representations highlighting the most informative features. The non-local block is shown in Figure 4.4. X^B is fed into three separate convolutions to embed the feature map. The non-local operation f, together with g, a simple linear embedding, computes the relationship between different locations. Then a residual connection is applied, followed by an average pooling to downsample feature maps to the size of $N \times 1024$.

For affective and visual features, because of the relatively small feature dimension, the self-attention mechanism is not applied to them. Instead, an average pooling (AP) layer is added, which works on the temporal dimension, to make affective and visual features have appropriate sizes with the behavioral features.

Finally, these three individual's features are concatenated to derive the refined individual feature map:

$$\mathbf{H} = \left[\alpha\left(X^B\right), \mathrm{AP}\left(X^A\right), \mathrm{AP}\left(X^V\right)\right]. \tag{4.3}$$

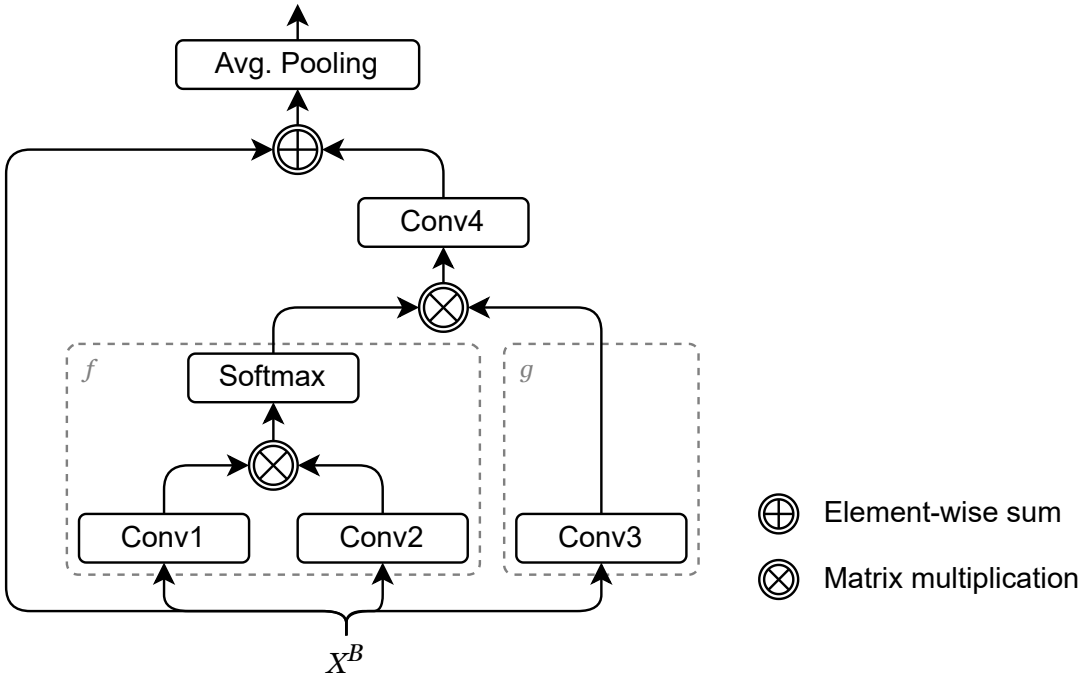

FIGURE 4.4: Self-attention block. The convolutional layers are all with a kernel size of $1 \times 1 \times 1$, but have different weights.

where α is the attention operation.

Note that in Equation (4.3), $\mathbf{H}$ is represented in terms of feature types. It is reorganized according to the detected people. Without ambiguity, the notation $\mathbf{H}$, *i.e.*, $\mathbf{H} = [h_1, ..., h_N]$ is still used, where each h_i contains behavioral, affective, and visual features obtained from individual learning, particularly h_1 is for the elderly, and $[\cdot, \cdot]$ denotes concatenation. This $\mathbf{H}$ is then used as the input to the next module: group learning.

4.2.3 Group Learning

EE of the elderly relies on subtle interactions among individuals present in a multi-party social HRI scenario. It has been known that estimating engagement solely from the elderly is not very reliable. In fact, in the elderly-nurse-robot interaction scenario, nurses are not just the auxiliaries and participants of the interaction. They are also the people who are in daily contact with the elderly and hence they

have the prompt judgment about the expressions of the elderly. These judgments are conveyed in their behaviors.

Generally in human conversation, each participant plays a specific role: speaker, addressee, or side-participant who is part of the group of potential speakers but is currently taking on a listening role [170, 171]. In a wild dynamic multi-party interaction, the main interaction group is defined to consist of participants, such as the speaker, addressee, and side-participants. The rest, who may be bystanders and overhearers, is called the background. In the elderly-nurse-robot interaction scenario described above, it should be hypothesized that *analyzing all participants in the main interaction group and their relationships help to estimate the engagement of the individual elderly.*

To represent the main interaction group, a graph structure is designed where each node corresponds to a participant and stores his/her feature map, and each edge represents the interaction between the participants of the two nodes. Graph neural networks (GNNs) [172] is bulit to learn the graph representation, which is to compute the hidden representation of each node in the graph by attending over the rest. Specifically, an adapted two-layer graph attention network (GAT) [173] is designed to learn the underlying interactions between nodes by computing attention weights for each edge. The input to the network is $\mathbf{H} = \{h_1, ..., h_N\}$ that are derived from individual learning. The network outputs a new set of transformed node features $\mathbf{L} = \{l_1, ..., l_N\}$.

First, following the approach of [173], a learnable transformation, which is parameterized by a shared weight matrix W, is applied to every node feature h_i in order to obtain a higher-level feature Wh_i.

The score e_{ij} of attention from node j to node i is computed by

$$e_{1j} = \mathbf{a}_1 \cdot Wh_1 + \mathbf{b}_1 \cdot Wh_j \tag{4.4}$$

$$e_{ij} = \mathbf{a}_2 \cdot Wh_i + \mathbf{b}_2 \cdot Wh_j, \quad \text{for } i \neq 1 \tag{4.5}$$

where $\mathbf{a}_j$ and $\mathbf{b}_j$ $(j = 1, 2)$ are the weight vectors to be learnt, and "$\cdot$" represents the dot product of vectors. Here $\mathbf{a}_1$ and $\mathbf{b}_1$ are for the attention from any node to node 1 only, and $\mathbf{a}_2$ and $\mathbf{b}_2$ are shared for all other situations. This special design is

due to the fact that the elderly are the main participant among all the participants in the group.

Next, a LeakyReLU nonlinearity is applied to the scores. They are further passed through a softmax operation to generate the normalized weights:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}(e_{ij})\right)}{\sum_{k=1}^{N} \exp\left(\text{LeakyReLU}(e_{ik})\right)}. \tag{4.6}$$

The normalized weights are used to compute the new node feature as a linear combination of the old features:

$$h_i' = \sum_{j=1}^{N} \alpha_{ij} W h_j. \tag{4.7}$$

To stabilize the learning process, multi-head structure is employed, where $K(>1)$ independent attention mechanisms execute the transformation, resulting in K different h_i'. Then, the new features of the i-th node $h_i^{k\prime}$ $(k = 1, \cdots, K)$ from every head are passed through the LeakyReLU activation function and concatenated to derive the updated node feature denoted by g_i:

$$g_i = \left[\text{LeakyReLU}\left(h_i^{1\prime}\right), \cdots, \text{LeakyReLU}\left(h_i^{K\prime}\right)\right]. \tag{4.8}$$

Here g_i is the output node feature of the first graph attention layer (GAL) and its dimension is K times that of h_i.

Finally, the updated node features $\mathbf{G} = \{g_1, ..., g_N\}$ is forwarded to the second GAL, which similarly goes through erefeq:e1j to erefeq:hiprime with h_j being replaced by g_j and with a new set of learnable weights W, $\mathbf{a}_i$ and $\mathbf{b}_i$ as well. This second GAL also takes K heads. Rather than concatenation, the node i of features $\mathbf{L}$ on this (prediction) layer is generated by averaging:

$$l_i = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{N} \alpha_{ij}^k W^k g_j \tag{4.9}$$

where α_{ij}^k are normalized weights computed for the k-th attention mechanism and

W^k is the corresponding weight matrix. Figure 4.5 illustrates this two-layer, K-head attention based group learning process.

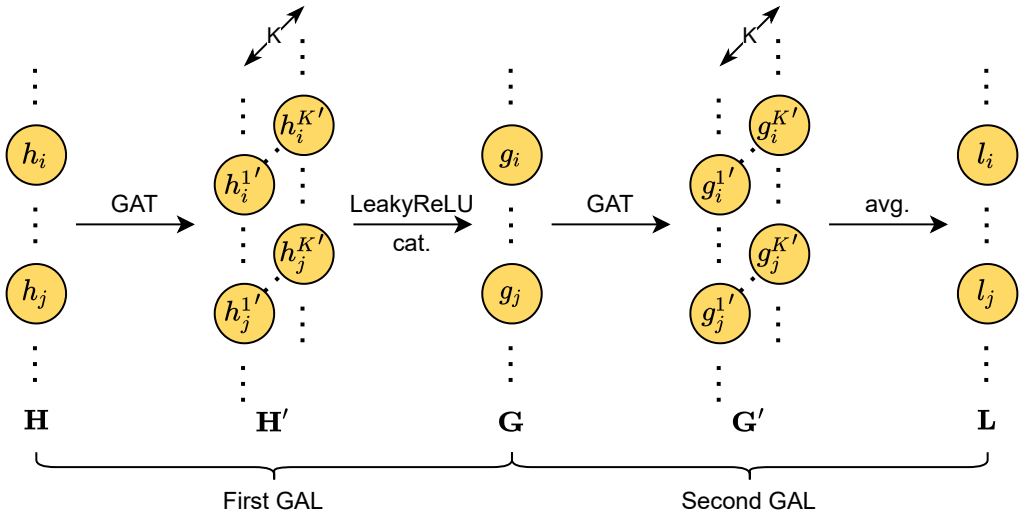

FIGURE 4.5: Schematic diagram of the group learning process with two layers and K-head attention mechanism.

4.3 Experiments and Results

4.3.1 BHEH Dataset

To the best of my knowledge, there is no publicly available labeled dataset for learning ERI, not to mention that in a multi-party scenario. In this study, the annotation of engagement level of the elderly are manually labelled based on a dataset called *BHEH* dataset, which Prof. Nadia M Thalmann shared with me.

BHEH dataset contains video recordings of real world ERI collected by Prof. Nadia M Thalmann and her colleagues from a nursing home in one of their research projects. Specifically, BHEH dataset records the elderly-robot interaction via a social humanoid robot, Nadine, with a human-like appearance. She was placed at the center of an elderly homeward. The interaction is recorded by five different cameras at different views. In the study, the data (video recordings) from the front view is exclusively utilized, *i.e.*, from Nadine's perspective containing most part of

the front face of the participants. The details of the data collection method can be found in [6, 174].

It should be noticed that there was no constraint imposed on the participants. The elderly talked to a socially intelligent robot while the nurses occasionally provided help. Moreover, the dataset was collected in a wild, dynamic, and multi-party environment. In the background, nurses might pass through the scene; some old people might sit near to or far from the interaction group; and the interaction participants might leave and join at any time.

In the experiments, 43 interaction sessions were manually annotated based on the Engagement of a Person with Dementia Scale. (EPWDS) [89]. The length of the videos is between 3 and 38 minutes (over 560 minutes in total). The number of participants for each session is from 1 to 6. The label of the data is the engagement score, which was obtained by normalizing the EPWDS engagement score to $[0, 1]$. Two components, which were deemed less relevant to the problem, were removed to simplify the process of EPWDS for artificial engagement annotation. The detail of the adapted annotation form is shown in Figure 4.6. Each interaction session was annotated at least by two experts. Figure 4.7 illustrates the labelled engagement statistics.

4.3.2 Implementation Details

For the original BHEH videos, a frame rate of 15 fps was used. To extract video features from the pre-trained networks, the videos were sampled by selecting one frame from every 5 frames, with a clip length of 32 frames. Consequently, each video clip captured an interaction period of approximately 10.67 seconds, which was employed as the input to the proposed model.

The dataset was randomly divided into 5 subsets, ensuring that no interaction session appeared in two sets. A 5-fold cross-validation was conducted to assess the model's performance. *i.e.*, one set as a testing set and the rest as a training set in every fold. The reported results are the average error.

Engagement Estimation Annotation Form								
Instruction This engagement estimation form contains three parts: ***Affective***, ***Visual***, and ***Behavioral***. For each video clip, you are expected to fill the table below (beginning & end timestamps and engagement value). The value indicates the extent to which you agree to the following statements for the elder person…								
Behavioral Engagement								
1					2	3	4	5
Responds to an activity by avoiding, shoving away, pulling back from, hitting, or mishandling the activity, the robot used, or the person/s involved.					...	Neutral	...	Responds to an activity by approaching, reaching out, touching, holding or handling the activity, the robot used, or the person/s involved.
Time Stamps								
Eng. Value								
Affective Engagement								
1					2	3	4	5
Displays negative affect such as apathy, anger, anxiety, fear, or sadness (e.g., disinterest, distressed, restlessness, repetitive rubbing of limbs or torso, frowning, crying).					...	Neutral	...	Displays positive affect such as pleasure, contentment or excitement (e.g., smiling, laughing, delight, joy, interest and/or enthusiasm).
Time Stamps								
Eng. Value								
Visual Engagement								
1					2	3	4	5
Appears inattentive, has an unfocused stare or turns head/eyes away from the activity, robot used, or the person/s involved.					...	Neutral	...	Maintains eye contact with the activity, robot used, or the person/s involved.
Time Stamps								
Eng. Value								

FIGURE 4.6: The simplified annotation form of EPWDS.

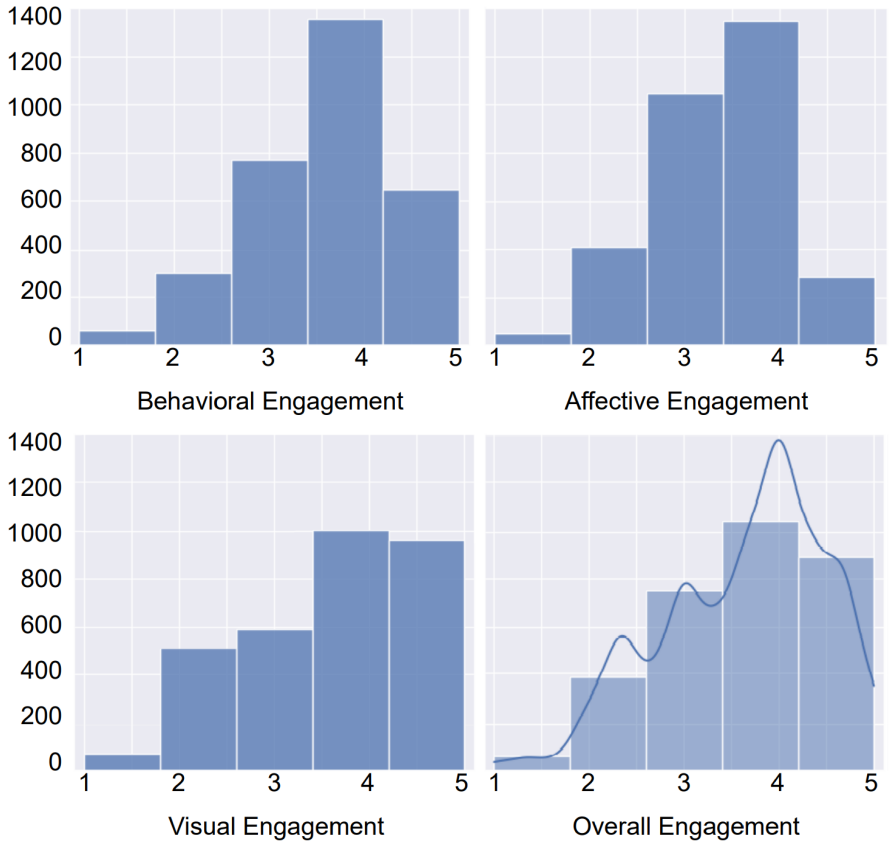

FIGURE 4.7: Overview of the engagement annotation. The horizontal axis and vertical axis represent the EPWDS engagement value and the video frame count, respectively.

In terms of the model, a pre-trained ResNet50-3D was utilized and adaptedas the backbone, followed by RoI Align to crop the behavior feature map into a size of 7×7. The self-attention module was applied to individuals' behavior representations by a non-local block with embedded Gaussian bottleneck. Affective and visual features were gained through DMUE [168] module and OpenFace [123] toolkit. The concatenated individuals' feature maps were fed into a two-layer, 3-head GATs module with a hidden size of 64, a dropout rate of 0.5, and slope $\alpha = 0.2$. The model was trained in two stages, beginning with the individual learning module and subsequently fine-tuning the network end-to-end, incorporating the adapted GAT module. Both stages were trained using the Adam optimizer in 80 epochs with an initial learning rate of 10^{-4}, divided by 10 every 40 epochs. The MSE loss is used in the training process.

4.3.3 Results

Although the proposed method should be compared with the state-of-the-art, there are no publicly available datasets or benchmarks for the target problem. The scenarios of HRI involved in the prior art are vastly dissimilar, posing challenges to conducting a fair comparison. As a compromise approach, the prior art methods are applied to the BHEH dataset. It should be emphasized that the codes for the prior art methods were not published and had to be developed from scratch.

For 2D CNNs [120], as the inputs of proposed approach were RGB videos without depth information, the version of using 2D body pose as features and AlexNet as model are selected to make comparison, which achieved similar performance reported in its paper. Inception V2 [116] had the same input format as the proposed method, so the implementation just followed its light-inception model architecture. In [122], LSTM was used to classify engagement based on extracted OpenFace and VGG features. To make fair comparisons, same frame length is used. For evaluation, two metrics are utilized, MSE of erefeq:mse and mean absolute error (MAE) defined as follows:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|y_i - \hat{y}_i| \tag{4.10}$$

where y_i is the predicted engagement value, $\hat{y}_i$ is the ground truth, and M is the number of sample clips. Note that MSE and MAE are chosen because they are two widely used criteria for visual regression tasks.

The results are reported in Table 4.2. The testing losses are shown in Figure 4.8. The performance of the proposed method (with MSE=0.0148) outperforms the prior art. For better understanding, three estimation examples is provided in Figure 4.9 which shows the center frame of the example clip, group detection, behavioral representation from body alignment, affective and visual representation, and estimation results from left column to right column. It can be seen that the proposed method achieves good results even under challenging conditions. Particularly, in examples 1 and 5, participants' bodies are detected and used for feature extraction, but the leftmost person is not desired. This is a counterexample of

the main group detection. In contrast, examples 2 and 3 detect all the participants successfully. In terms of affective and visual engagement representation, some inconsistency and instability occur due to the masks and senescent faces. For instance, the elderly in example 1 could not make meaningful expressions and the visual features in example 3 are also missed. This may explain why those methods only involving facial information often fail to produce good results. In addition, the elderly from examples 1 and 3 is not good at body language, so the information from side participants helps in the estimation, *e.g.*, the body representation captures the raised hand of the nurse.

TABLE 4.2: Engagement estimation of the elderly.

	MSE	**MAE**
Random Guess	0.1763	0.3435
2D CNNs [120]	0.1427	0.3711
Inception V2 [116]	0.0283	0.1649
LSTM [122]	0.1148	0.3030
Proposed Method	**0.0148**	**0.0996**

4.3.4 Ablation Studies

A number of ablations are conducted to analyze the proposed method. The results are reported in Table 4.3. $\mathcal{B}$, $\mathcal{A}$, and $\mathcal{V}$ are the results of using a single engagement element of behavioral, affective, and visual. It can be seen that the results are inferior to that produced by the proposed multi-element method. The self-attention module also helps improve the performance by 0.0105 in MSE and 0.0386 in MAE. By employing the GATs in group learning module, the results have 0.0148 increase in MSE, which means that the signals from side-participants contribute to the estimation in multi-party ERI. For the comparison of ordinary GAL and adapted GAL, the results show that the adapted model achieves better results, although this performance improvement is not very significant. Figure 4.10 illustrates the MSE and MAE results of these two kinds of graph layers.

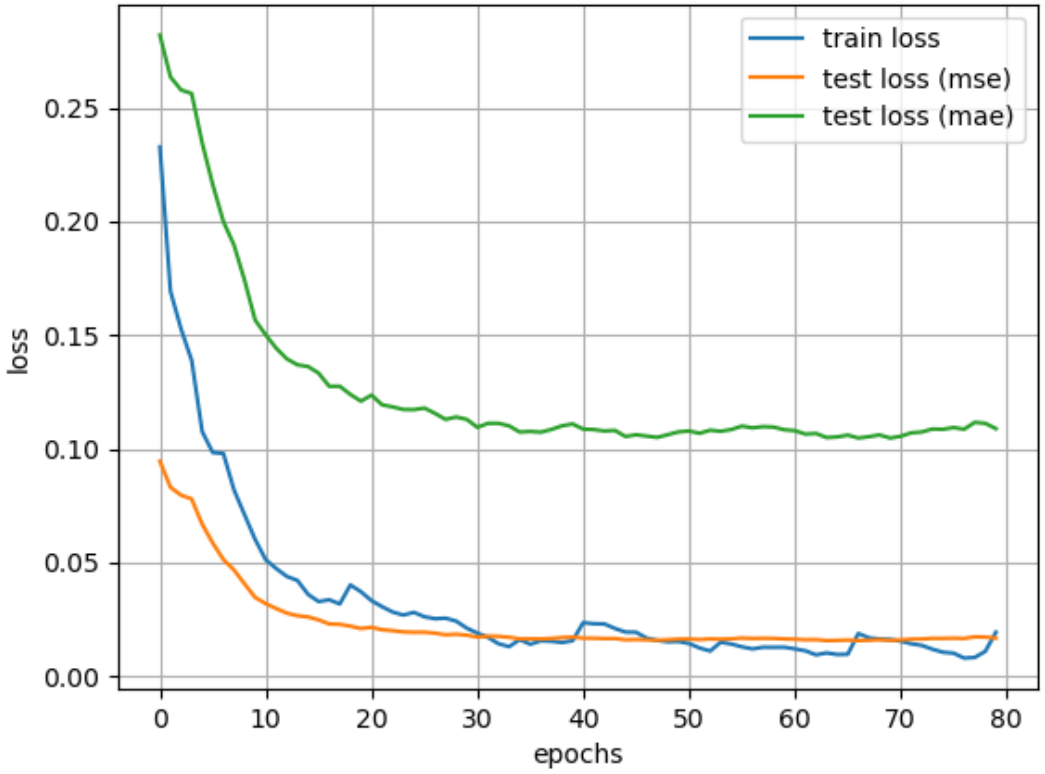

FIGURE 4.8: Losses of MSE and MAE on the testing set.

TABLE 4.3: Ablation results.

	MSE	**MAE**
$\mathcal{B}$	0.0451 (↓0.0303)	0.1750 (↓0.0754)
$\mathcal{A}$	0.1235 (↓0.1087)	0.3690 (↓0.2694)
$\mathcal{V}$	0.1567 (↓0.1419)	0.4184 (↓0.3188)
w/o self-attention	0.0253 (↓0.0105)	0.1382 (↓0.0386)
w/o GATs	0.0296 (↓0.0148)	0.1380 (↓0.0384)
original GATs	0.0173 (↓0.0025)	0.1080 (↓0.0084)

4.4 Conclusion

This chapter introduces an automatic approach for analyzing wild multi-party social HRI and estimating the engagement state of the elderly–the main participant– in the HRI. The proposed method involves the adaptation of pre-trained models to extract behavioral, affective, and visual features of the participants in the main interaction group from real-world videos of such interactions. To predict the engagement state of the elderly, a deep learning model is constructed, comprising a

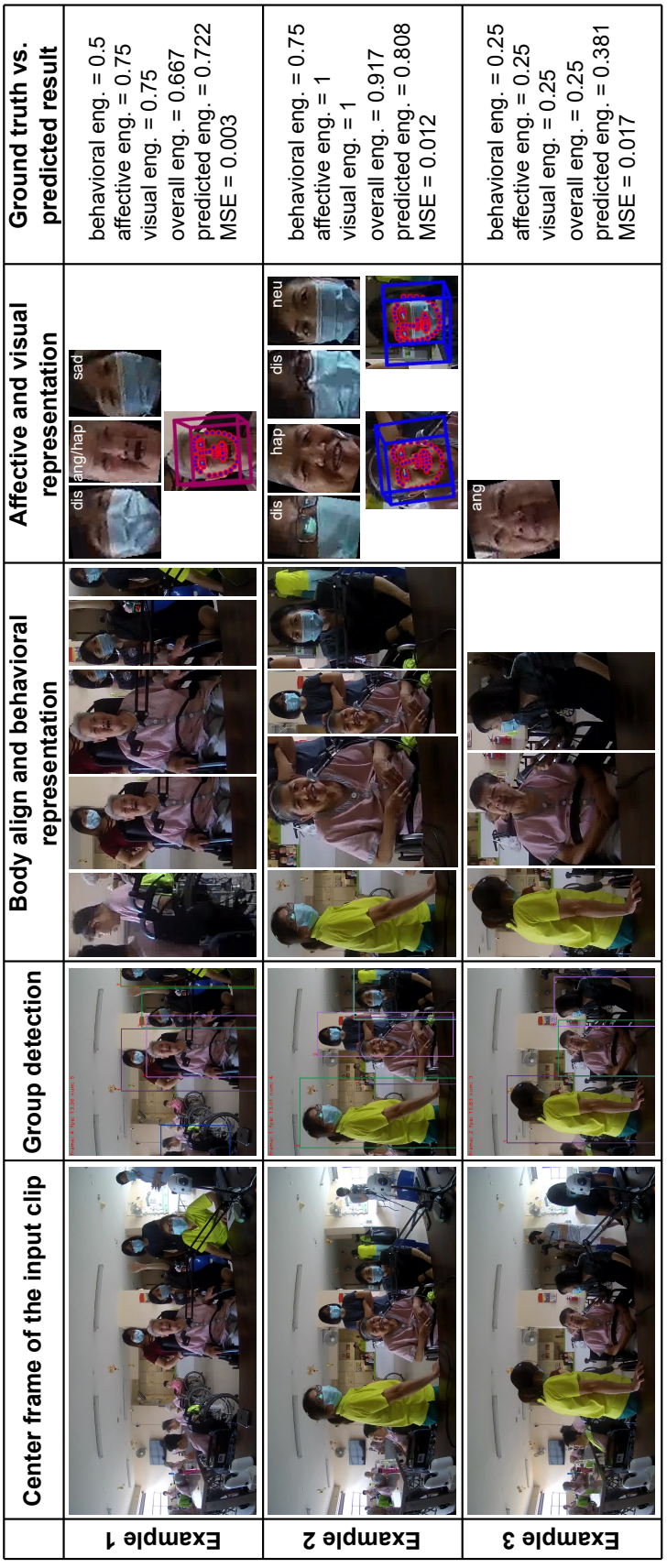

FIGURE 4.9: Visualization of the engagement estimation results.

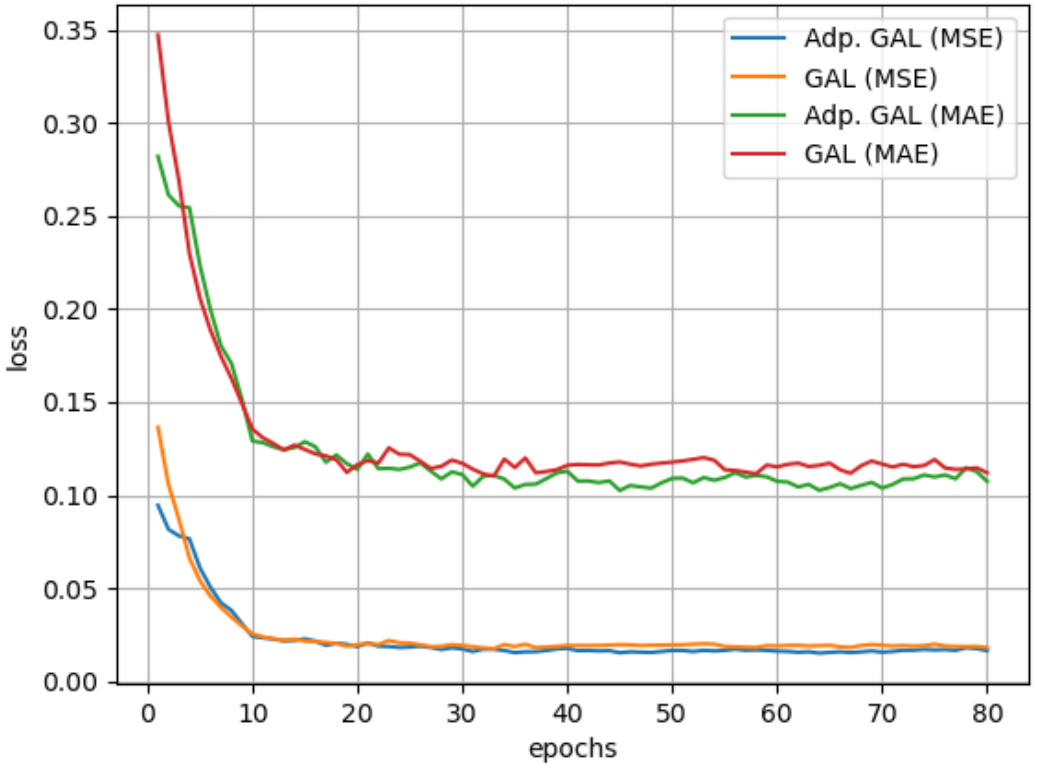

FIGURE 4.10: Comparison between ordinary GAL and the adapted GAL in terms of the MSE and MAE losses.

self-attention network for individual learning and a graph attention network for group learning. This model takes into account the multi-modal features of all participants as input and estimates the elderly's engagement level. Also, a labeled engagement dataset was created. By leveraging multi-modal features and incorporating individual and group learning mechanisms, the proposed method effectively predicts engagement and demonstrates superior performance compared to existing approaches, as evidenced by the experimental results.

Chapter 5

Personality Estimation and Emotion Recognition in Social Interaction

5.1 Introduction

Personality and emotions play an important role in social interactions and have impacts on the cognition of human behavior. Previous research shows that people's judgments about others' personality and emotions will affect their feeling and making decisions. An example is that in a job interview, the information conveyed by the interviewee through non-verbal signals will affect his/her interview results. Hence automatic estimation and recognition of personality and emotions have various applications such as for job interview analysis and designing social robots.

This chapter considers the problem of estimating personality and emotions in human social interactions. In the last decades, computer vision techniques have achieved great progress in human pose estimation, action recognition, and facial expression analysis. Taking advantage of this, many emotion recognition methods have been introduced and developed. While some works attempt to utilize contextual information to improve the accuracy of emotion recognition, many approaches

focus on the development of robust and fast models based on facial features. For the scenarios with more than one person, the recognition of each person's emotion and the interplay between emotions are rarely studied.

Also, relatively less attention has been paid to personality estimation where the main task is to estimate the first impression of a single person [175]. First impression, also known as apparent personality, is the assessment results from peers, as opposed to real personality (self-reported). Existing methods for estimating apparent personality mainly analyze the behavior and expression of a single person from self-introduction videos. When it comes to multi-person scenarios, *i.e.*, face-to-face interactions in daily life, it becomes necessary to consider context information because people's actions and expressions are determined not only by their personality but also by the external environment and the interlocutor(s) [134].

In this chapter, a deep learning network to simultaneously estimate personality and recognize emotion in social interaction scenarios is presented, using non-verbal social signals and sociodemographic information from participants. The focus is on answering whether information from the interlocutor and estimated personality contributes to emotion recognition. The key hypotheses are: (i) personality estimation and emotion recognition both analyze the non-verbal social signals, which means that the solutions for these two tasks are closely related; (ii) context information is beneficial to the target tasks in interaction scenarios; and (iii) personality, as a factor for emotion generation, should be taken into consideration in the task of emotion recognition. These hypotheses will be validated on two datasets, which will be discussed later.

The overall pipeline of the proposed approach is shown in Figure 5.1. The input is multi-view social interaction videos and sociodemographic information. After extracting body and face image sequences in the extractor module, body learning and face learning modules are employed to gain high-level representations, where a multi-branch ResNet-Attention network [176, 177] is utilized to process the non-verbal social signals of the target and interlocutor(s) on temporal-spatial dimensions. Then, the personality of the target person is estimated based on

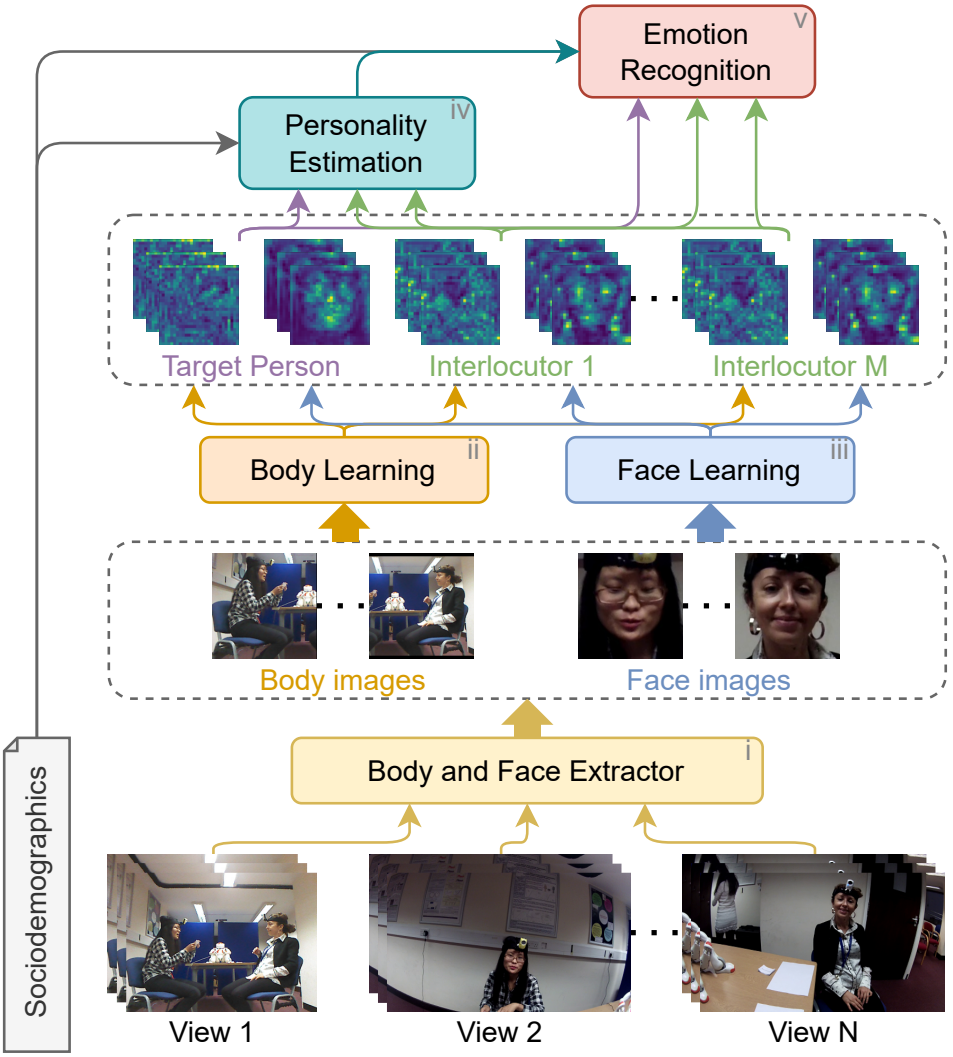

FIGURE 5.1: Overview of the proposed context-aware and personality-based emotion recognition approach.

these features. The emotion is finally recognized based on the learned body, facial, sociodemographic features as well as the estimated personality. The main contributions of this chapter include

- A novel approach is proposed to jointly estimate personality and recognize the emotions of participants in social interaction scenarios.
- Different from the prior art, the newly proposed method employs the same architecture to analyze personality and emotions, which in addition is capable of estimating both apparent and real personality.

- The social interaction context is emphasized and utilized. The multi-modal data from the target individual and interlocutor(s) are utilized, forming a context-aware structure to estimate personality, and then use personality to improve the accuracy of emotion recognition.
- A set of experiments have been conducted to examine the impact of information from the target individual and interlocutor(s) on personality estimation, as well as the effects of personality on emotion recognition.

5.2 Preliminaries

Before starting to describe the proposed approach to personality assessment and emotion recognition, this section briefly discusses some pertinent concepts that inspired us to design an architecture like the current one.

5.2.1 Non-verbal Behavior

Research has shown that non-verbal behavior plays a very important role in people's social lives. The generation of human non-verbal behavior is influenced by many factors, *e.g.*, personal determinants, conscious or unconscious goals, environments, and perceptual and affective information [58, 61, 62]. Personal determinants include biology, culture, gender, and personality; perceptual and affective information are the outcomes of processing interaction and interlocutor(s). Taking multi-party interaction as an example (Figure 5.2), four participants form an interaction. Background information and personal determinants are the high-level factors, where determinants influence the decision-making, perceptual and cognitive-affective processes. At the same time, the latter two interact with each other, and influence a person's behaviors together with the determinants.

Context-aware and personality-based emotion recognition, referring to the model above, includes context and personality as inputs. Context-aware refers to the awareness of the context in human-human social interaction. In this chapter, the context includes the interlocutors' behaviors and interaction settings, unlike

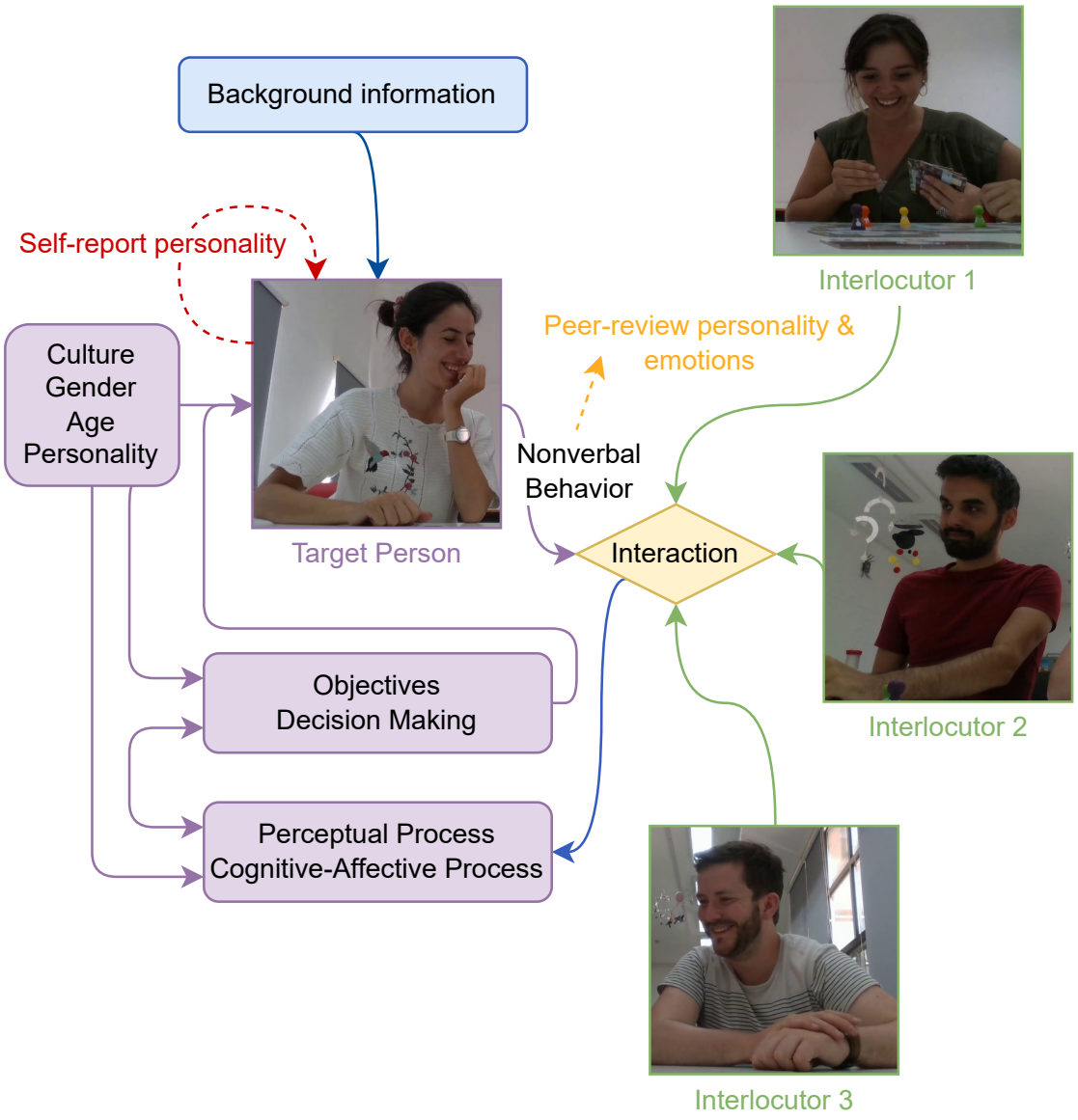

FIGURE 5.2: Illustration of social interaction and the factors that influence people's behavior. The dot arrows denote personality and emotion assessment processes, where yellow represents the generation of apparent personality as well as emotions, and the red one represents the generation of real personality.

the previously mentioned context-aware emotion recognition approaches. Certain scenarios, such as indoor conversations and game playing, are only considered. Therefore, in this work, background environment is relatively fixed, so the proposed model does not take the surrounding background into account. From the previous theoretical model, it is known that emotions are influenced by the context and further react in behavior. At the same time, personality influences the

perceptual and cognitive-affective processing of the context, which also acts on the generation of behavior. Theoretically, context contributes to personality estimation and emotion recognition; context and personality help with emotion recognition.

5.2.2 Personality and Emotions

Personality is used to describe a person's character that reflects a set of behavior, emotion, and cognition patterns. To quantify personality, traits are used to provide multidimensional aspects of personality. Several trait theories exist, such as Big-Five Inventory (BFI) [91] and HEXACO [92], *etc*. In this chapter, either BFI or HEXACO will be used according to the datasets. The specific description of the personality traits can be found in Section 2.2.3.

In psychological science, researchers have proposed many different theories of emotion models, such as basic emotions and dimensional models. In this chapter, category emotion theories are used to represent participants' emotions provided by the datasets, which includes two sets of emotion labels: Set A = {positive, small positive, small negative, negative} and Set B = {anxious/frustrated/angry, bored, confused, delighted}.

5.3 Proposed Approach

The proposed method takes multi-view human-human social interaction videos $V = \{v_1, v_2, ..., v_N\}$ as input, where N denotes the number of views. Given this, the target personality $\hat{P} \in \mathbb{R}^{d_p}$ and emotion $\hat{E} \in \mathbb{R}^{d_e}$ are estimated. d_p and d_e are the dimensions of personality traits and emotion classes.

To achieve this goal, a context-aware and personality-based deep learning architecture is designed, which learns the representations of the body and facial information from the target person and interlocutor(s). This model mainly consists of five modules, as shown in Figure 5.1, (i) body and face extractor, (ii) body learning, (iii) face learning, (iv) personality estimation, and (v) emotion recognition modules.

5.3.1 Body and Face Extractor

At the very beginning, the responsibility of body and face extractor $\mathcal{E}$ is to preprocess the raw multi-view video data gaining model-friendly and the most important body and face image sequences, X^B and X^F. Because the intention is to analyze the non-verbal signals from the body and face in social interaction scenarios, interaction-independent background information is not relevant to the problem and even negatively interferes with the experimental results. Thus, the body and face extractor can be represented as

$$\mathcal{E} : V \rightarrow \{X^B, X^F\}. \tag{5.1}$$

Here, $X^B = \{X^{TB}, X^{IB_1}, ..., X^{IB_M}\}$, where TB and IB denote target body and interlocutor(s) body, M is the number of interlocutor(s). In addition, since X^{TB} and X^{IB} have the same data structure and similar subsequent processing methods, their representations are simplified to X^b when no distinction is necessary. X^F has a similar data structure and X^f denotes X^{TF} or X^{IF}.

The multi-view videos are divided into two categories: ego-view and general-view videos. Ego-view videos are captured by a camera fixed on the participant's head, which is used to record the visual information from participants' perspectives. In contrast, general-view videos are captured by a camera placed next to the interaction group, which is used to capture the overall scene including almost the full bodies. However, it should be noted that clear and complete face or body pictures do not always exist, because of head turning and body movement during the interaction. In addition, since the data in experiments are multi-view videos, it is necessary to extract the bodies and faces of the participants in a way that ensures (a) the image sequences from different views are synchronized as much as possible and (b) the acquired body and face image sequences are optimal, *i.e.*, the clearest and most complete faced and bodies are selected among different views.

First of all, due to the inconsistent frame rates of the ego-view and general-view recordings, body and face images obtained from two data sources cannot always align in time. To solve this problem, the ego-view videos are sampled using the frame rate distribution of the general-view videos, because the frame rates of the

scene videos sometimes are lower and unstable. More specific analysis of the data can be found in Section 5.4. Moreover, the videos are sampled with a stride of 5 to reduce the computational cost and to increase data intra-variation. Finally, the videos are cut into non-overlapping clips with a chunk size of L.

Then a body detection algorithm is implemented to find the body bounding boxes of participants in each frame of the general-view clips. In frames where pedestrians appear, they are removed based on their proportion of appearance in the entire session. In addition, to keep the input image sequences spatially stable and prevent body displacement caused by different bounding box positions, for each person, a maximum bounding box is calculated to cut the person out of the entire session video. Subsequently, while keeping the aspect ratio constant, the cropped body images are resized and normalized into $\mathbb{R}^{L\times H\times W\times 3}$, in which $H \times W$ is the image size. The blank pixels are filled with black.

After comparing facial information in videos from different views, the optimal video source is selected to extract facial information, *i.e.*, the one including clearer frontal facial information. Specifically, from the body image gained from the previous step, face are detected and cropped. Then, detected face images are aligned and resized to the same size as body images. As mentioned earlier, because of the unconstrained interaction scenarios, sometimes the captured images are blurred or even completely lost. In order to remedy this, nearest-neighbor interpolation is employed to fill in the missing frames with face photos that can be found before or after. Specifically, within the time period $[t_i, t_j]$ when face detection fails, the faces that can be detected by t_{i-1} are used to fill the interval $[t_i, ..., t_{(i+j)/2}]$, and faces from t_{j+1} are used to approximate facial features in $[t_{(i+j)/2+1}, ..., t_j]$. To some extent, this interpolation indeed brings undesired temporal mismatches, but, under the settings discussed in this chapter, it can be assumed that in face-to-face interactions, the emotions and facial expressions of the participants will not change rapidly, and the experimental data also is basically in line with this assumption.

5.3.2 Body and Face Learning Modules

Given $\{X^B, X^F\}$ being the pair of synchronized image sequences, X^b and $X^f \in \mathbb{R}^{L\times H\times W\times 3}$, a multi-branch ResNet-attention based deep learning architecture is designed to learn the high-level spatio-temporal representations of body and face that used to predict the personality and emotion of a target person. As Figure 5.3 illustrates, four separate branches TB, TF, IB, IF are designed. TB and IB are responsible for learning the body representations Z^B from target person and interlocutor in body learning module $\mathcal{B}$, whereas TF and IF learn the facial features Z^F in face learning module $\mathcal{F}$.

Body learning module ($\mathcal{B}$). The body learning module is composed of four functions: a backbone network (BB), a self-attentive mechanism (A), an average pooling layer (AP), and a fully connected layer (FC). Specifically, the backbone network is responsible for extracting the personality and emotion-related body behavior patterns in temporal and spatial domains. Thanks to the impressive progress in deep video understanding network [160, 178], ResNet-3D-50 [160] is used, which has competitive performance for action recognition, to capture the body features. The final residual block is pruned from the original network to keep more information and higher feature map resolution. Let chunk size L being 32, the model learns a feature map with size $4 \times 14 \times 14 \times 1024$ for every body image sequence X^b.

Then, a self-attention mechanism is employed to explore the important information in spatial and temporal dimensions, and refine the feature representations. This mechanism enables the model to focus on keyframes, such as the expressive moment, and local body regions. A non-local block [169] is used as the self-attention block, which calculates the response at any given position as a weighted sum of the features at all positions. The optimized features have the same dimensions as the previous ones. Subsequently, an average pooling is performed to reduce the feature size to 1×1024, followed by a fully connected layer that learns a projection from 1024 to 512 making the body representation fitting for later concatenation. The entire body learning module can be written as

$$\mathcal{B} : Z^B = \mathrm{FC}\left(\mathrm{AP}\left(\mathrm{A}\left(\mathrm{BB}\left(X^B\right)\right)\right)\right), \tag{5.2}$$

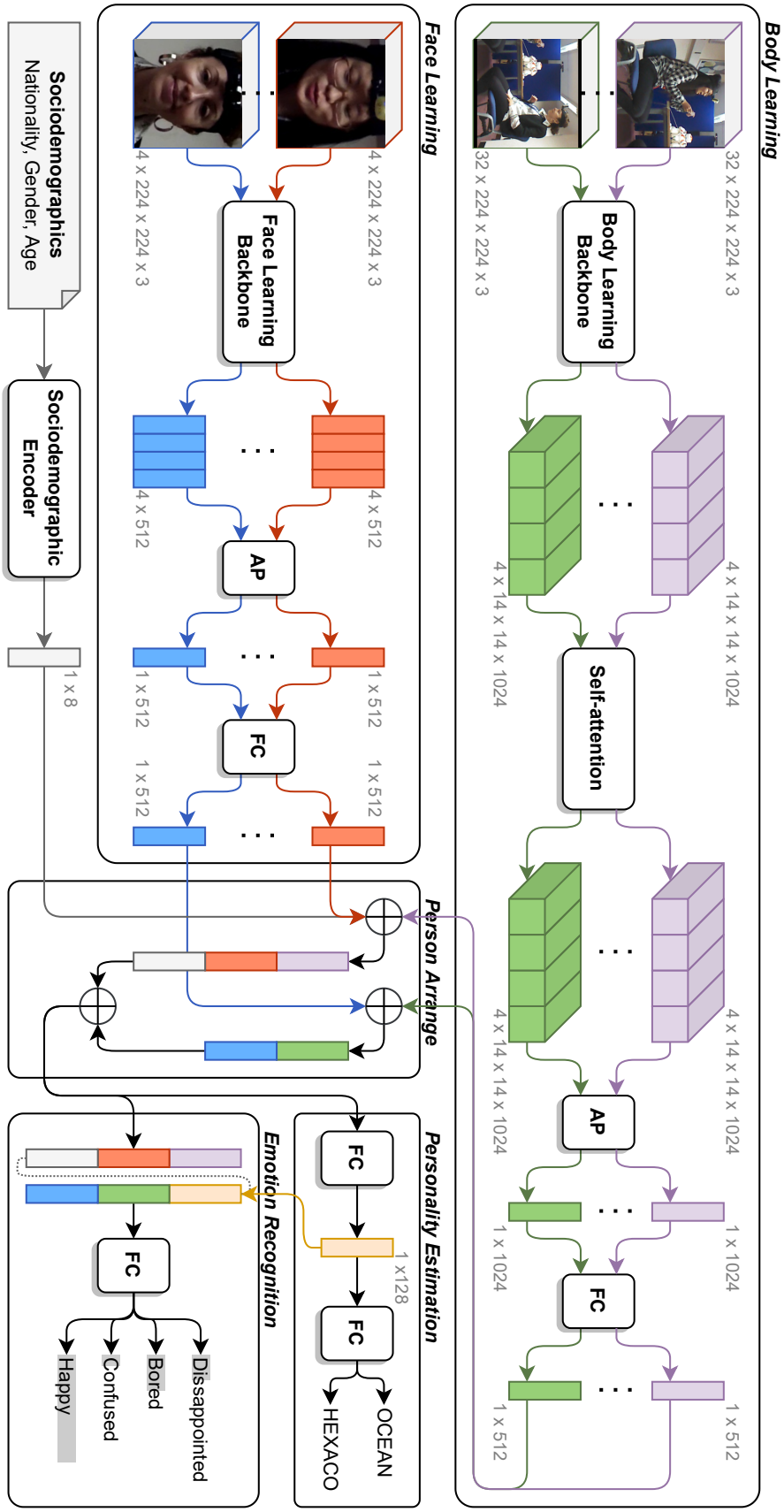

FIGURE 5.3: Architecture of the proposed context-aware and personality-based emotion recognition model.

where $Z^B = \{Z^{TB}, Z^{IB_1}, ..., Z^{IB_M}\}$. Similarly, $Z^b \in \mathbb{R}^{1\times1024}$ is used to represent their general form.

Face learning module ($\mathcal{F}$). In the face learning module, ResNet-18 [176] is used as the facial backbone (FB) for facial feature extraction. Here, each of the 4 frames are fed into the network to obtain a 4×512 dimensional feature matrix. These four frames are uniformly sampled from the original 32-frame chunk. Also, inflated 3D network is not employed to process facial image sequences, because facial expressions are usually presented naturally and the information in the temporal dimension is relatively unimportant. Therefore, using a 2D CNN network can effectively reduce the number of parameters. Next, the extracted facial features are forwarded to an adapted emotion recognition block, namely latent distribution mining and pairwise uncertainty estimation (DMUE) [168] by removing the final output layer. Finally, a average pooling layer is conducted for the facial feature map on the temporal domain to obtain a facial feature vector with a size of 1×512, *i.e.*, a feature representation of the same size as the output of body learning:

$$\mathcal{F} : Z^F = \mathrm{FC}\left(\mathrm{AP}\left(\mathrm{FB}\left(X^F\right)\right)\right), \tag{5.3}$$

in which $Z^F = \{Z^{TF}, Z^{IF_1}, ..., Z^{IF_M}\}$. The backbone model details are listed in Table 5.1.

5.3.3 Personality Estimation and Emotion Recognition

Before personality estimation and emotion recognition can be performed, two steps need to be completed in advance. One of them is the encoding of sociodemographics. In the proposed framework, sociodemographic information together with the social signals from interlocutor(s) constitutes the contextual information, which is defined as context-aware structure. In this work, the information used includes cultural background, gender, and age. These factors are encoded separately to obtain 6, 1, and 1-dimensional vectors, respectively. Cultural background is represented as the one-hot vector of 6 nations; gender is either 0 or 1; age is the normalized value between 0 and 1. Their concatenation forms the sociodemographic feature

TABLE 5.1: The structure of the backbone networks in body and face learning modules.

Layer Name	Body Learning	Face Learning
conv1	$5 \times 7 \times 7$, stride 2, 2, 2	7×7, stride 2, 2
maxpool1	$2 \times 3 \times 3$, stride 2, 2, 2	2×2, stride 2, 2
res2	$\begin{bmatrix} 3 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
maxpool2	$2 \times 1 \times 1$, stride 2, 1, 1	N.A.
res3	$\begin{bmatrix} 3 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
res4	$\begin{bmatrix} 3 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
res5	pruned	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$

$S^T \in \mathbb{R}^{1\times 8}$. It is important to note that only information from the target person is used.

So far, the body, facial, and sociodemographic information are acquired, and the next step is to integrate these features for the final inference. To achieve this, previously learned features are regrouped according to the relevant participants, as illustrated in Figure 5.3. Basically, the person arrange module is introduced to enable the learned representations to serve as input for subsequent personality estimation and emotion recognition modules in the order of the interaction participants, *i.e.*, body-and-face-level features are organized into individual-level features. Mathematically, this procedure can be written as

$$\{Z^B, Z^F, S^T\} \rightarrow \{Z^T, Z^{I_1}, ..., Z^{I_M}\}, \tag{5.4}$$

where $Z^T \in \mathbb{R}^{1\times 1032} = Z^{TB} \oplus Z^{TF} \oplus S^T$ is the feature from target person; $\oplus$ is the concatenation operation. $Z^{I_{(\cdot)}} \in \mathbb{R}^{1\times 1024}$ is the obtained feature(s) from interlocutor(s).

Personality Estimation ($\mathcal{PE}$). For the personality estimation module, the estimation task is treated as a regression problem, because the experimental results of the regression are more informative than classification, as described in Section 5.4. A simple two-layer fully connected neural network is utilized. The extracted features from the target person and other participants are selectively concatenated and fed into the network with the hidden size of 128, *i.e.*, the feature is projected into the dimension of 128 and then reduced to 1. The model separately estimates different personality traits.

Here, the datasets contain personality labels in the form of OCEAN or HEXACO. The overall estimation is

$$\mathcal{PE} : \hat{p} = \mathrm{FC}\left(\mathrm{FC}\left(Z^{T}, Z^{I_1}, ..., Z^{I_M}\right)\right). \tag{5.5}$$

The output $\hat{p}$ represents the estimated score of a particular trait of real or apparent personality. The mean squared error loss $\mathcal{L}_{\mathcal{P}}$ is used to train personality estimation module

$$\mathcal{L}_{\mathcal{P}} = \frac{1}{K}\sum_{i=1}^{K}\left(p_i - \hat{p}_i\right)^2, \tag{5.6}$$

where K denotes the mini-batch size and p_i is the ground truth personality score.

Emotion Recognition ($\mathcal{ER}$). The last step is to use the extracted features from the target person and interlocutor(s), as well as the estimated personality to recognize emotions. The final features $Z = \{Z^{T} \oplus Z^{P} \oplus Z^{I_1} \oplus ... \oplus Z^{I_M}\}$, in which $Z^{P} \in \mathbb{R}^{128}$, being the representation of personality, is the output from the first FC layer in personality module.

In contrast to personality estimation, a classification model is used because the data are labeled by the category emotion theory. Again, a two-layer fully connected neural network is used, which takes the obtained feature representations as input, and project them into a 128-dimensional space. The final step is to calculate the probabilities of 4 emotion categories with a Softmax function, *i.e.*,

$$\mathcal{ER} : \hat{e} = \sigma\left(\mathrm{FC}\left(\mathrm{FC}\left(Z\right)\right)\right), \tag{5.7}$$

where σ denotes the Softmax operation. Let $c = [1, ..., C]$ denoting the emotion category, w_c being the weight of class c, e_c and $\hat{e}_c$ representing the ground truth label and inferred probability of class c respectively, the lost function of emotion recognition module is designed as a weighted cross-entropy loss

$$\mathcal{L}_{\mathcal{E}} = \frac{1}{K} \sum_{i=1}^{K} \left(- \sum_{c=1}^{C} w_c \cdot e_{i,c} \log \hat{e}_{i,c} \right). \tag{5.8}$$

Therefore, the total training loss is defined as $\mathcal{L} = \lambda_P \mathcal{L}_{\mathcal{P}} + \lambda_E \mathcal{L}_{\mathcal{E}}$, where λ_P and λ_E are the weights for personality estimation and emotion recognition losses.

5.4 Experiments and Results

5.4.1 Datasets

Personality estimation and emotion recognition have attracted attention from researchers recently, so there are some publicly available datasets. Unfortunately, most datasets address only one of these two tasks, but the objective of the target problem is to infer the emotions that arise in social interaction, which is far from typical scenarios and has its own unique character. Therefore, some popular datasets are not suitable for this task, such as EMOTIC [143] and CAER [144].

On the other hand, MHHRI [112] collects videos of dyadic human-human interaction, accompanied by the real personality reported by the interlocutor themselves as well as the apparent personality evaluated by their peers, which is suitable for us to explore the influence of context on personality estimation, and meanwhile, the difference between real and apparent personality can be analyzed. MUMBAI [179] records four-player board game interactions via multiple cameras. The dataset is annotated with emotional moments, personality, and game experience. Therefore, MHHRI and MUMBAI are chosen to evaluate the proposed personality estimation and emotion recognition modules. The examples of two datasets are visualized in Figure 5.4.

FIGURE 5.4: Examples of multi-view social interaction datasets. The first row, from left to right, shows frames of the general-view, ego-view 1, and ego-view 2 videos in MHHRI. The second row includes two general-view frames in MUMBAI.

Particularly, MHHRI includes 12 social interaction sessions with a total of 18 participants, where sessions last around 10 to 15 minutes. The raw videos are provided as 20 to 120-second clips, so they have to be rearranged to shorter clips containing $L = 32$ frames, which have the same size as the network input. MUMBAI includes 43 sessions of 4-player board game interaction. The video recording duration is approximately 4 to 22 minutes.

In terms of personality labels, MHHRI uses BFI-10 to measure the OCEAN personality traits. As mentioned before, BFI-10 consists of 10 phrases and each will be discretely assessed as 1 to 10, where two items correspond to one personality trait. By averaging and normalization, the ground-truth labels between 0 and 1 are gained. In addition, self-report and peer-review annotations represent the real and apparent personality, respectively. In contrast, MUMBAI is labeled using the peer-review HEXACO. The same processing procedure is conducted to gain the normalized labels. The distributions of personality labels are shown in Figure 5.5. For the emotion classification task, two sets of emotion labels provided by MUMBAI are used, *i.e.*, emotion set A includes {positive, small positive, small negative, negative}, and set B consists of {anxious/frustrated/angry, bored, confused, delighted}. It should be noticed that positive emotions are more common

than negative ones.

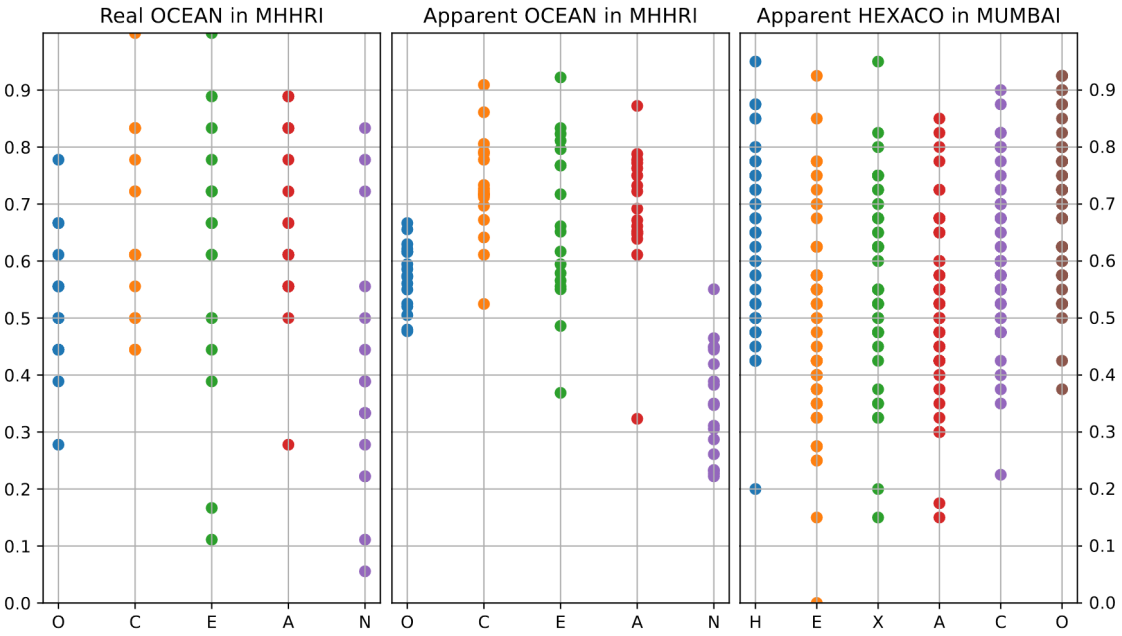

FIGURE 5.5: Visualization of the normalized ground truth distributions of OCEAN and HEXACO personality on MHHRI and MUMBAI datasets.

5.4.2 Implementation and Evaluation Metrics

In order to obtain the body and face image sequences, YOLOX [180] and OpenFace [123] are used, as the extraction tools, to find the body bounding boxes of all participants from the general-view videos, and to detect, crop and align the faces from ego-view videos of MHHRI and from general-view videos of MUMBAI. Before sending the body and face image sequences to the network, a set of data augmentation approaches are performed to get more diverse inputs, *e.g.*, random affine transforms (translation, scaling, and rotation), horizontal flip, and random resized crops with a probability of 0.5. The training of the body backbone involves pre-training on Kinetics [164]. For the face learning module, pretrained ResNet-18 networks on AffectNet [181] are utilized to extract and learn facial features. The reason to use pretrained models is to improve the robustness of estimation, as the datasets used in experiments are relatively small.

For the MHHRI dataset, frames from ego-view videos are selected according to the frame distribution of the general-view videos. As shown in Figure 5.6, the general-view videos sometimes have lower and fluctuated frame rates. It can be seen that their frame rates are from 5 to 30 fps, shown as green dots. The blue dots are some examples of general-view video clips showing very unstable frame intervals.

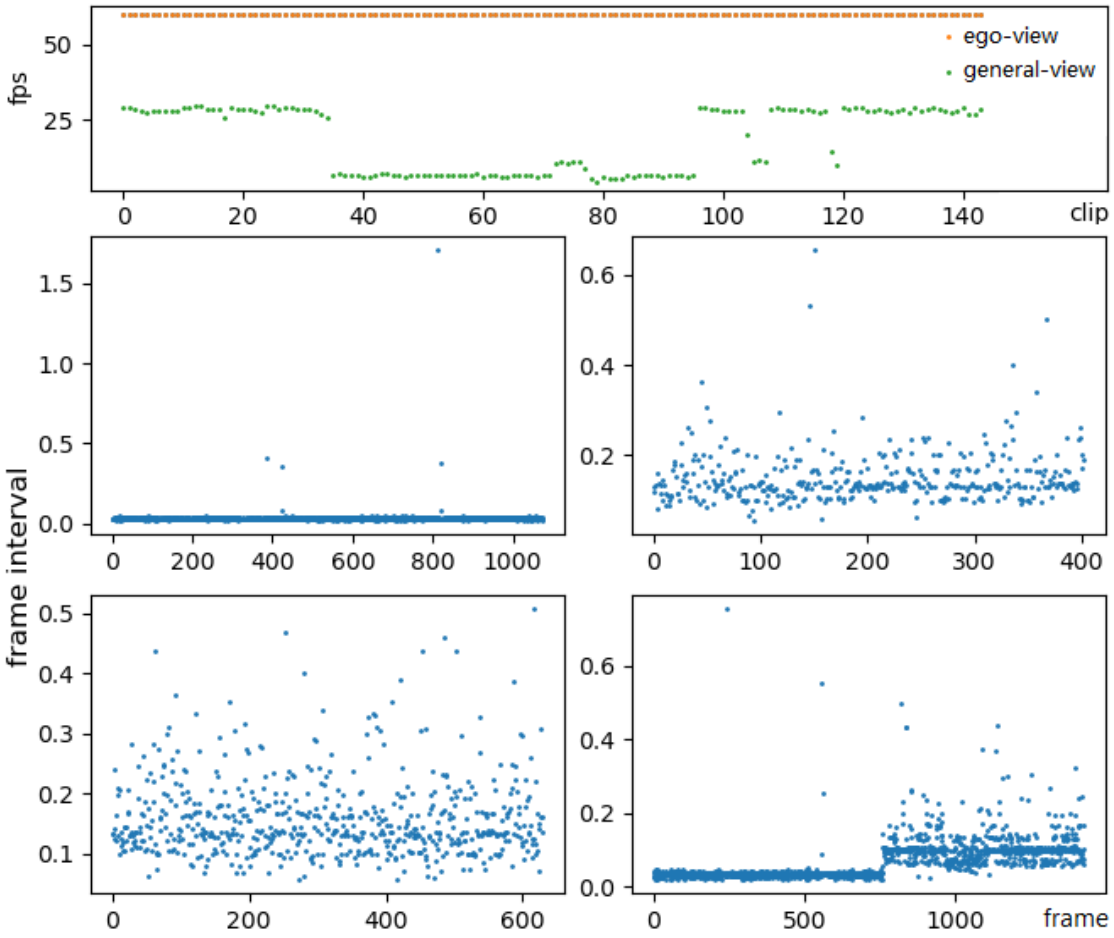

FIGURE 5.6: A glance at the multi-view videos. The top figure illustrates the frame per second of ego-view and general-view videos. The bottom figures show the frame interval of four sample clips in seconds.

The input size of the face and body images is $H \times W = 224 \times 224$ pixels. The network takes $L = 32$ frames as input. Finally, the input image sequences are about 1 to 7 seconds for MHHRI and 5 seconds for MUMBAI. The datasets are then randomly divided into 6 subsets to create a 6-fold validation. In both MHHRI and MUMBAI, there are people who participate in more than one interaction session. Therefore, the separation of datasets is completed under the condition

that there is no individual appearing in both sets. The model is trained using the Adam optimizer with an initial learning rate of $lr = 1e{-}4$ and a mini-batch size of $bs = 4$. As the datasets used are relatively small, MHHRI is trained for only 1 epoch, while MUMBAI is trained for 10 epochs. The weight values for personality estimation λ_P and emotion recognition λ_E are both set to 0.5.

To examine the effects of various features, the simplest architecture is used initially and more features are gradually introduced. The personality estimation performance is evaluated using the target person's body features (TB) and his/her facial features (TF), individually. Then, the body and facial features (TB-TF) are combined to form features all about the target person. Next, the context information is added, *i.e.*, the body features of the target and interlocutors (TB-IB), the face features of the target and interlocutor (TF-IF), and all body and face information from all participants (All). The results of a non-end-to-end method using body and facial information retrieved by the pretrained models are also presented. Finally, the comparison between proposed approach and state-of-the-art is provided. To evaluate the performance of emotion recognition, comparison with baseline models is conducted while investigating the effect of personality.

To evaluate the effect of the proposed personality estimation approach, the metrics in previous personality estimation works are used [126, 130, 131]. The evaluation aims to compute the accuracy of all personality traits among all input video clips, *i.e.*,

$$Acc = 1 - \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \,, \tag{5.9}$$

where N is the number of input samples; $\hat{y}_i$ is the estimated trait score; y_i is the ground truth. In addition, the R-squared is calculated, which is defined as

$$R^2 = 1 - \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \,. \tag{5.10}$$

Acc and R^2 are then averaged over five traits to evaluate the overall performance. The final Acc and R^2 are between 0 and 1 from bad to good performance.

As for the personality estimation, the problem is also considered as a binary classification task. Those ground truth less than 0.5 is classified as negative samples, on

the contrary, greater than or equal to 0.5 is treated as positive samples. Classification results are evaluated by the accuracy *Acc* and F_1 score. The evaluation of the emotion recognition approach has a similar evaluation metric F_1 score but in the form of the average of each class with weighting parameters, as well as precision and recall.

5.4.3 Personality Estimation Results

The personality regression results of MHHRI and MUMBAI are shown in Table 5.2 and Table 5.3, respectively. Table 5.4 illustrates the classification results of MHHRI compared with existing works. Also, the visualization of the body and face learning and internal feature maps are demonstrated in Figure 5.7.

Comparison of body and face features. Overall, facial information is more useful than body information in estimating personality, whether it is real or apparent. The reason might be the properties of the data, *i.e.*, the interactions in MHHRI and MUMBAI adhere to particular patterns. Participants in MHHRI sit face-to-face and talk with each other, whereas participants in MUMBAI play board games, periodically moving their arms to carry out game-related actions. As a result, body behaviors are generally straightforward and consistent. As shown in Table 5.2 and Table 5.3, when estimating the real personality based on the target individual's face (TF) and body (TB) information, the results of using facial features performs better (0.865/0.832) than the results using body features (0.853/0.817), achieved on both MHHRI and MUMBAI. The results employing both the facial features (TF-IF) and body features (TB-IB) of two participants lead to a similar conclusion. This advantage is more pronounced in the estimation of apparent personality, where accuracy increases by 0.01.

Effect of information from the interlocutor. Comparing the estimation accuracy of TB *vs*. TB-IB, TF *vs*. TF-IF, and TB-TF *vs*. All, it could be found that, in real personality estimation, adding features from the interlocutor(s) helps to improve the model performance, for example, 0.853 *vs*. 0.865 and 0.865 *vs*. 0.867 from adding IB to TB and adding IF to TF, respectively. [134] came to a similar conclusion, that openness, extraversion, and neuroticism traits gained benefits by

TABLE 5.2: The regression results of real and apparent personality prediction on MHHRI.

	Methods *	O		C		E		A		N		Avg.	
		Acc	R^2	*Acc*	R^2	*Acc*	R^2	*Acc*	R^2	*Acc*	R^2	*Acc*	R^2
Real	Baseline	0.737	0.904	0.704	0.872	0.686	0.852	0.691	0.857	0.698	0.865	0.703	0.870
	All (F.)	0.759	0.925	0.745	0.927	0.720	0.908	0.807	0.931	0.779	0.932	0.762	0.924
	TB	0.888	0.980	0.862	0.970	0.845	0.948	0.850	0.963	0.817	0.955	0.853	0.963
	TF	0.904	0.987	**0.887**	0.976	0.852	0.966	0.852	0.959	0.826	0.949	0.865	0.967
	TB-TF	0.902	0.986	0.862	0.974	0.845	0.958	0.851	0.962	0.819	0.955	0.856	0.967
	TB-IB	0.904	0.983	0.883	0.977	0.838	0.958	0.858	0.968	**0.842**	0.954	0.865	0.968
	TF-IF	**0.912**	0.986	0.880	0.974	0.847	0.958	0.858	0.966	0.840	0.959	0.867	0.968
	All	0.907	0.984	0.875	0.975	**0.860**	0.964	**0.867**	0.972	0.829	0.946	**0.868**	0.968
Apparent	Baseline	0.739	0.907	0.689	0.856	0.707	0.873	0.704	0.872	0.720	0.888	0.712	0.879
	All (F.)	0.731	0.922	0.634	0.860	0.680	0.875	0.651	0.870	0.819	0.961	0.703	0.898
	TB	0.925	0.993	0.937	0.991	0.870	0.976	0.923	0.988	0.938	0.994	0.919	0.988
	TF	**0.960**	0.998	**0.938**	0.991	0.898	0.984	0.930	0.988	0.944	0.994	**0.934**	0.991
	TB-TF	0.945	0.996	0.929	0.989	0.879	0.981	0.932	0.988	**0.944**	0.995	0.926	0.990
	TB-IB	0.921	0.992	0.929	0.990	0.887	0.981	**0.938**	0.992	0.927	0.993	0.920	0.990
	TF-IF	0.954	0.997	0.931	0.989	**0.899**	0.984	0.927	0.989	0.944	0.995	0.931	0.991
	All	0.927	0.993	0.937	0.987	0.875	0.978	0.935	0.990	0.936	0.993	0.922	0.988

* Baseline method adopt here is the random guess; All (F.) denotes the method using all four branches but not trained end-to-end, *i.e.*, only the extracted features from pre-trained models are fed into the net; TB, TF, TB-TF, TB-IB, TF-IF, and All denotes the approaches with activated branch name(s).

TABLE 5.3: The regression results of the real personality estimation on MUMBAI.

Methods	**H**		**E**		**X**		**A**		**C**		**O**		**Avg.**	
	Acc	R^2	*Acc*	R^2	*Acc*	R^2	*Acc*	R^2	*Acc*	R^2	*Acc*	R^2	*Acc*	R^2
TB	0.839	0.851	0.713	0.720	0.787	0.795	0.840	0.855	0.837	0.850	0.888	0.917	0.817	0.831
TF	0.845	0.861	0.710	0.717	0.847	0.880	0.843	0.858	0.850	0.891	0.899	0.944	0.832	0.859
TB-TF	0.845	0.861	0.724	0.733	0.844	0.878	0.851	0.860	0.844	0.881	0.897	0.944	0.834	0.860
TB-IB	0.844	0.861	0.715	0.724	0.801	0.812	0.844	0.858	0.841	0.879	0.891	0.920	0.823	0.842
TF-IF	**0.854**	0.872	0.717	0.728	**0.859**	0.901	0.854	0.862	**0.861**	0.903	0.904	0.946	**0.842**	0.869
All	0.849	0.866	**0.730**	0.741	0.842	0.879	**0.867**	0.887	0.858	0.899	**0.907**	0.948	**0.842**	0.870

TABLE 5.4: Comparison of personality classification results with benchmark models.

	Methods	O		C		E		A		N		Avg.	
		Acc	F_1	*Acc*	F_1	*Acc*	F_1	*Acc*	F_1	*Acc*	F_1	*Acc*	F_1
Real	FPV-HMS+SPV-HOG [112]	-	0.66	-	0.54	-	0.54	-	**0.61**	-	**0.57**	-	0.58
	Proposed (All)	0.93	**0.82**	0.85	**0.70**	0.84	**0.61**	0.71	0.40	0.69	0.51	0.80	**0.61**
Apparent	CNN+LSTM [132]*	0.40	0.30	0.44	0.33	0.33	0.07	0.50	0.60	0.41	0.58	0.42	0.38
	VGG DAN+ [131]*	0.48	0.19	0.62	0.48	0.54	0.15	0.44	0.55	0.46	0.66	0.51	0.41
	3DCNN [129]*	0.46	0.26	0.50	0.39	0.39	0.06	0.38	0.53	0.43	0.65	0.43	0.38
	3DResNet [130]*	0.46	0.22	0.62	0.55	0.43	0.03	0.41	0.49	0.54	0.71	0.49	0.40
	SPV-HOG [112]	-	0.67	-	0.60	-	0.61	-	0.67	-	0.65	-	0.64
	Proposed (TF)	**0.85**	**0.67**	**1.00**	**1.00**	**0.82**	**0.71**	**0.96**	**0.91**	**0.99**	**0.92**	**0.92**	**0.84**

* Original methods are proposed in [129–132] and implemented by [128] on MHHRI dataset.

adding context features and the surrounding environment. On the other hand, only IB adds positive effects to apparent personality estimation. Unexpectedly, when adding the interlocutor's facial features (IF) or both facial and body features (IF-IB), the results do not improve. Here, the results from [112] demonstrate that multi-modal features boost the performance for extroversion, which is aligned with results reported here, in terms of TB-IB and TF-IF. Overall, the features of interlocutor are helpful for real personality prediction, but in apparent prediction, this effect is small, and sometimes even unfavorable.

Estimation of real and apparent personality. Generally, the regression performance of apparent personality is much better than the results of real personality. Comparing the best average results in these two tasks, real personality estimation gets 0.8676 in accuracy, whereas apparent personality estimation achieves 0.9338. The distribution of ground truth labels might provide some cues. As shown in Figure 5.5, the labels of the apparent personality are more concentrated with smaller variance, which means that estimation is easier to give a value close to the real value. On the contrary, extraversion and neuroticism in real personality are with larger variance, resulting in the most challenging tasks. Also, [112] has made a Pearson correlation analysis on peer-reviewed and self-report personality showing that significant correlations only can be found for extroversion and conscientiousness. Moreover, psychological studies found that self-peer agreement correlations tended to rise as the number of peer raters increased [182, 183]. Therefore, there is a correlation between appearance and real personality, but this convergence is not significant when the sample size is small. This is why the same architecture is used but separately trained.

Regression *vs*. classification. Classification experiments were also conducted and the results are listed in Table 5.4. It should be emphasized that this work focuses on predicting personality using a regression approach, so a new model was not trained using the classification-related loss function, but rather intuitively dividing the outputs into two parts, *i.e.*, above or below 0.5, based on the regression results. Except for the experimental results of this model, other data are provided by [112, 128]. A same dataset splitting method in [128] is used, while [112] uses double leave-one-subject-out cross validation. As for ground truth, [112, 128] partitioned the data according to the mean value. Therefore, the comparison of

classification results here is not very fair, but perhaps it can provide some value of reference. As shown, the proposed approaches outperform other existing methods.

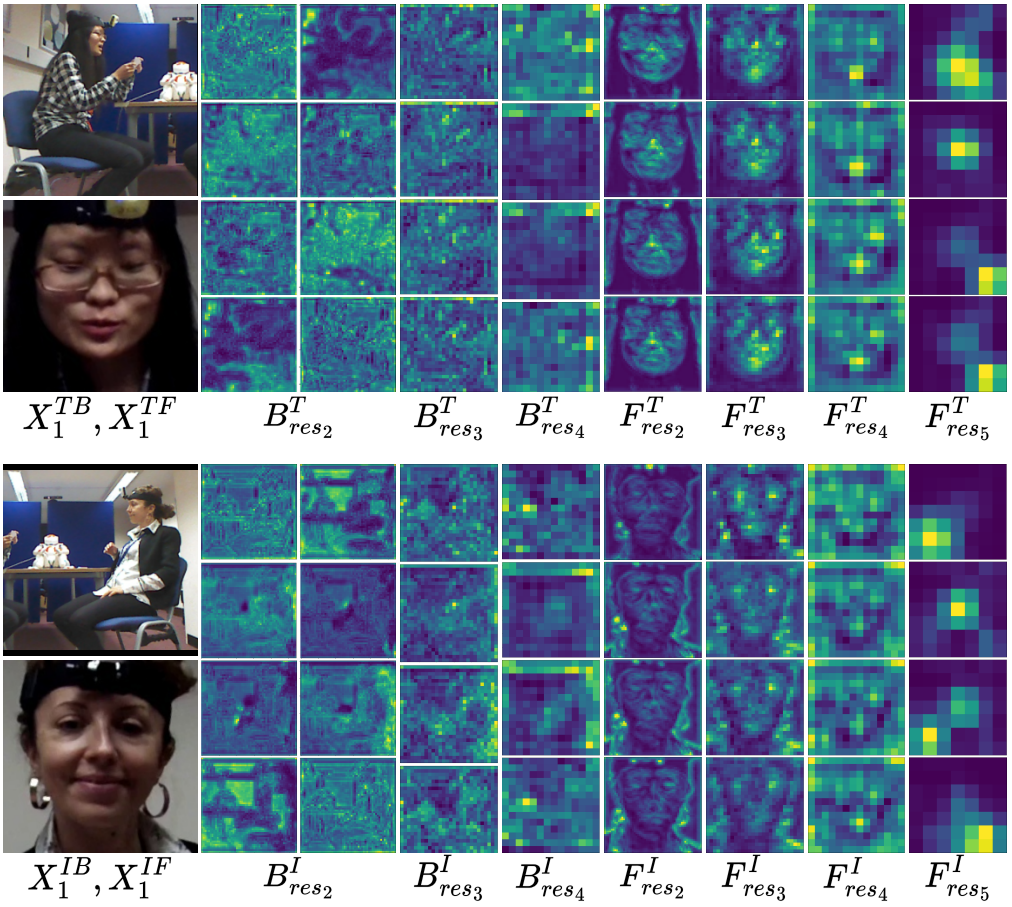

FIGURE 5.7: Visualization results of the body and face learning and internal feature maps from the target (T) and interlocutor (I). For each person, the first column shows the initial frame of body and face image sequences, followed by body (B) and face (F) feature maps from corresponding layers.

5.4.4 Emotion Recognition Results

Emotion recognition results on MUMBAI are shown in Table 5.5. Comparison is made between the proposed approach and the baseline models in [179]. It can be seen that the proposed model outperforms them.

Personality-based *vs*. non-personality-based. According to the experimental results shown in Table 5.5, using the predicted personality as one of the features

to evaluate emotions has better performance. This is also consistent with the hypothesis that the estimated personality information can improve emotion recognition performance. It can be seen that, in two different emotion annotations, the evaluation results (F_1, Precision, and Recall) achieved by proposed method are significantly improved. Among them, the results of the method without using personality as a feature achieved $F_1 = 0.631$ in set A, which is 0.164 higher than the best score in the baseline. When trained with personality estimation and using the predicted personality to guide emotion recognition, the performance of the model is further improved, achieving 0.642 in terms of F_1 score. A similar effect can be found in set B.

TABLE 5.5: Emotion recognition results on emotion annotation set A and set B from the MUMBAI dataset.

	Methods	F_1	Precision	Recall
Set A	KNN w/ face high	0.359	0.406	0.359
	DT w/ face high	0.380	0.390	0.392
	RF w/ face	0.467	0.463	0.482
	ELM w/ face high	0.431	0.456	0.415
	LSTM w/ face	0.453	0.442	0.490
	Random	0.218	0.250	0.248
	Proposed w/o pers.	0.631	0.620	0.636
	Proposed w/ pers.	**0.642**	**0.674**	**0.719**
Set B	KNN w/ face high	0.272	0.297	0.277
	DT w/ face all	0.282	0.292	0.282
	RF w/ face low	0.316	0.366	0.311
	ELM w/ face all	0.293	0.307	0.299
	LSTM w/ face all	0.309	0.308	0.323
	All delighted	0.213	0.186	0.250
	Proposed w/o pers.	0.594	0.633	0.617
	Proposed w/ pers.	**0.615**	**0.703**	**0.690**

5.5 Conclusion

This chapter presents a deep learning approach that simultaneously estimates personality traits and recognizes emotions in social interaction scenarios. The proposed method utilizes non-verbal social signals from the target person and the

interlocutor(s) to construct a context-aware structure. The predicted personality traits are then incorporated to develop a personality-guided architecture for emotion recognition. A series of experiments are conducted to analyze the influence of various feature modalities, contexts, and personality traits on emotion recognition. These experiments include comparisons with existing approaches as well as ablation studies. The results demonstrate that contextual features play a significant role in accurately predicting real personality traits, while information from the context and the target person's personality contribute to improved emotion recognition performance. The findings indicate that the proposed method achieves good performance for both personality estimation and emotion recognition tasks.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In conclusion, multi-party social HRI is a challenging and less explored research area. To design a socially intelligent agent capable of handling multi-party conversations, various issues have to be addressed. The overall objective of this thesis is to develop new techniques to empower IAs with the ability to understand the behaviors, intentions, and affective states of the participants in multi-party social interaction. In general, the goals of the project have been met by proposing solutions for recognizing and interpreting participants' behaviors, intentions, and affective states in multi-party social HRI by analyzing non-verbal social signals, including engagement intention, engagement during interaction, personality, and emotions. These solutions are expected to equip IAs with the ability to enhance the user experience in multi-party social HRI scenarios. The experiments and analyses conducted show that the use of multi-modal features outperforms using a single modality and that social information from other participants contributes significantly to the estimation of engagement intention, engagement, personality, and emotion.

The method of estimating engagement intention before multi-party social HRI is proposed first. In this task, the literature on social behavior science is reviewed to

select promising non-verbal social cues. After this, a novel approach to estimating human engagement intention in multi-party social HRI is presented leveraging multi-modal information, including image features and social signals. Then, necessary features are extracted, generated, and categorized. The method is built on the CNN-LSTM network, which takes image features and social signals as input, making use of general information conveyed in images, semantic social cues proven by social psychology studies, and temporal information in the sequence of inputs. Also, the proposed network has a multi-branch structure, allowing for different types of inputs. A novel feature transition method is designed to interpret multi-party social signals. Finally, the method is evaluated on three datasets, and the experimental results show that the proposed method can infer human engagement intention well. Moreover, using multi-modal features improves the estimation of engagement intention.

When an interaction starts, the goal of engagement estimation is to infer the inner state of a participant attributing to being together with the other participants and continuing the interaction. Moreover, the use of social robots in healthcare systems or nursing homes to assist the elderly and their caregivers will be becoming common. A supervised machine learning approach is presented for estimating the engagement state of the elderly in multi-party human-robot interaction scenarios using real-world video recordings. The method is based on the fundamental concept of engagement components in geriatric psychiatry and HRI video representations. It adapts pre-trained models to extract behavior, affective, and visual signals to form multi-modal features. These features are then fed into a neural network made of a self-attention mechanism and average pooling for individual learning, a graph attention network for group learning, and a fully connected layer to estimate engagement level. The proposed method is evaluated using 43 wild multi-party elderly robot interaction videos. The experimental results show that the proposed method is capable of detecting the key participants and estimating the engagement state of the elderly effectively. Also, the signals from side participants in the main interaction group considerably contribute to the EE of the elderly in the multi-party ERI.

Finally, personality and emotion are investigated in multi-party social interaction, which have great influences on people's cognition and behavior. There are

fewer works on personality estimation. The contribution of personality to emotion recognition has also not been adequately studied. Therefore, a context-aware deep learning framework has been proposed, which automatically estimates the personality of a target person in human-human social interaction scenarios, based on the target person's own and the interlocutor's body behavioral and facial information. Then, this network is expanded to form a context-aware and personality-based emotion recognition framework, *i.e.*, first estimating the personality and then recognizing the emotions based on the estimated personality. A set of experiments have been conducted showing that the proposed method has good performance in both personality and emotion experiments.

6.2 Future Work

The research on multi-party human-robot social interaction presented in this thesis focuses on the analysis of non-verbal social signals from the interlocutors and potential interactors, *i.e.*, assessing their engagement, personality, emotions, *etc.* However, the linguistic information of the interaction participants and robotic behavioral planning have not been sufficiently discussed. Meanwhile, non-verbal behavior analysis as part of IAs' perception and processing system, together with natural language processing outcomes, can guide the generation of behavioral responses and verbal language, which are some possible and interesting future research directions. Some of the directions are given below.

Emotion and non-verbal behavior generation in multi-party social HRI. In this thesis, the processing of non-verbal behavior refers to the perception and understanding of human participants in multi-party social human-robot interaction. If these processes were to be mapped to analogous functions in the human brain, they would correspond to perceptual, cognitive, and affective processes. However, it is not enough for an intelligent agent, as a socially intelligent entity, to simply process the information it receives. The generation of emotion and the optimization of emotion-based non-verbal behaviors are two important tasks that

are necessary to endow social robots with social attributes. Only when the design of the closed-loop interaction is completed, can the robot contribute to the interaction in a manner similar human beings.

In multi-party social HRI, emotions and non-verbal behaviors are affected by many factors, as discussed in Chapter 2. However, in multi-party scenarios, taking emotions as an example, the analysis of emotions is more complicated, as group-based emotions should be considered. Group-based emotions are based on individuals' self-categorization as a group member and occur in response to situations perceived as relevant for that group [184]. Figure 6.1 and Figure 6.2 show widely accepted group emotion models and the emotion generation process.

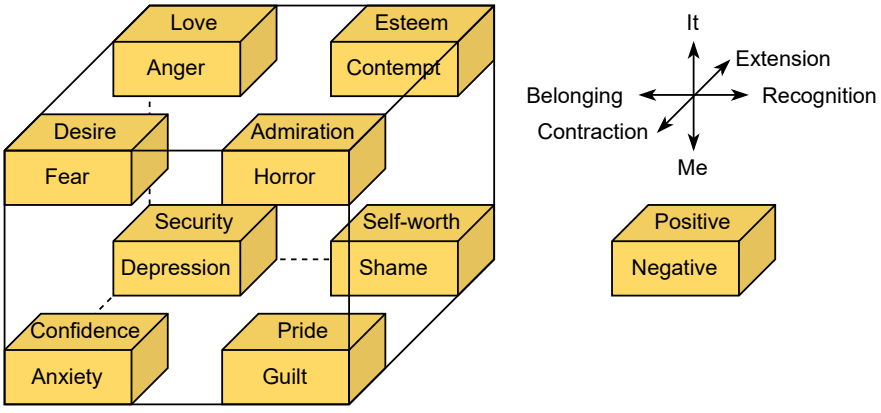

FIGURE 6.1: Group-level emotion model defined by a matrix of four interpersonal factors.

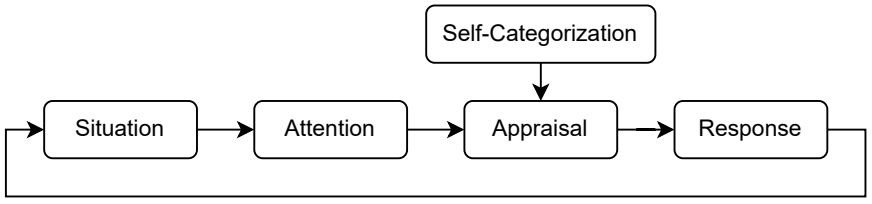

FIGURE 6.2: Group-level emotion generation process.

Robot memory for multi-party social HRI. In the future, the memory manager of the IAs, solving experience-related problems, can provide more personalized solutions for multi-party human-robot social interaction by generating responses that are more in line with human experience. Specifically, in socially intelligent agents, the memory manager is responsible for encoding, storing, and retrieving

past experience, which is important for improving user engagement and supporting cognitive capabilities such as reasoning and decision making [185].

For example, for a multi-party engagement estimation task, a socially intelligent agent should make different inferences based on the memory of each interaction participant and potential participants. In human society, people may have different behavior if they take different roles, *i.e.*, social behavior towards an acquaintance or a stranger is usually different. In addition, the memory function is also helpful to estimate and tracking engagement in a personalized way. People from different cultural or social backgrounds may have various social actions, not to mention their personalities. Therefore, the systematic approach can address ordinary problems, but only with memory, the social agents can make more human-like perceptual and cognitive processes. However, there are many challenges to adding a memory component. Considering the data for learning the behavior of an individual, the amount must be enormous. Therefore, an ideal method is to divide the learning process into holistic and individual ones, where the former is used to make general estimations and the latter for individual adaption based on memory.

Multi-party social dialog management. Several approaches for engagement estimation and affective-related estimation tasks to ad- dressing dynamic human-robot conversations have been proposed in this thesis. Nevertheless, many studies claim that the designing of a multi-party dialog system is the key to multi-party social companions because socially intelligent agents should have the ability to talk with multiple people in a small group. Dialog management inherently is a huge topic involving natural language processing, dialog system design [186], speaker diarization [187], turn-taking management [188], addressee and response selection [189], *etc*. In addition, although current research arts in the above areas resolve different issues separately, these topics have a lot of overlapping areas.

Moreover, memory also plays a key role in dealing with dynamic conversations that are common in multi-party social HRI. Considering dyadic interaction, the conversational dynamics are not obvious. However, in multi-party social HRI, more than one conversation can be freely suspended or activated and participants can join or leave at any time. To converse within a group, robots must overcome several challenges. Therefore, memory, in this situation, enables the social agents to manage

these dynamics, *i.e.*, social agents can reason about past conversations by, recalling previous memories. The capacity to recall information from previous conversations helps understand concurrent conversations where topics and participants in each of the conversations and sub-conversations are recorded.

List of Publications

Journal Articles

- **Zhijie Zhang**, Jianmin Zheng, and Nadia Magnenat Thalmann, "Engagement estimation of the elderly from wild multiparty human-robot interaction," *Computer Animation and Virtual Worlds (CAVW)*, 2022, pp. e2120.
- **Zhijie Zhang**, Jianmin Zheng, and Nadia Magnenat Thalmann, "Context-Aware Personality Estimation and Emotion Recognition in Social Interaction," *The Visual Computer (TVCJ)*, 2023.

Conference Proceedings

- **Zhijie Zhang**, Jianmin Zheng, and Nadia Magnenat Thalmann, "Engagement Intention Estimation in Multiparty Human-Robot Interaction," in *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 2021, pp. 117–122.
- **Zhijie Zhang**, Jianmin Zheng, and Nadia Magnenat Thalmann, "Real and Apparent Personality Prediction in Human-Human Interaction," in *International Conference on Cyberworlds (CW)*, 2022, pp. 187–194.

Bibliography

[1] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in hri: An exploration of social and task-based context," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 2012, pp. 421–428. xv, 14, 15, 29

[2] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi, "Pepper Learns Together with Children: Development of an Educational Application," in *Proceedings of the IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 270–275. xv, 14, 15

[3] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, "Improving social skills in children with ASD using a long-term, in-home social robot," *Science Robotics*, vol. 3, no. 21, p. eaat7544, 2018. xv, 14, 15

[4] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, "Modeling Engagement in Long-Term, in-Home Socially Assistive Robot Interventions for Children with Autism Spectrum Disorders," *Science Robotics*, vol. 5, no. 39, p. eaaz3791, 2020. xv, 14, 15, 60

[5] W. Moyle, C. J. Jones, J. E. Murfield, L. Thalib, E. R. A. Beattie, D. K. H. Shum, S. T. O'Dwyer, M. C. Mervin, and B. M. Draper, "Use of a Robotic Seal as a Therapeutic Tool to Improve Dementia Symptoms: A Cluster-Randomized Controlled Trial," *Journal of the American Medical Directors Association*, vol. 18, no. 9, pp. 766–773, 2017. xv, 14, 25

[6] G. Tulsulkar, N. Mishra, N. M. Thalmann, H. E. Lim, M. P. Lee, and S. K. Cheng, "Can a humanoid social robot stimulate the interactivity of cognitively impaired elderly? a thorough study based on computer vision methods," *The Visual Computer*, vol. 37, pp. 3019–3038, 2021. xv, 14, 15, 73

[7] P. Ekman, "Universal Emotions," nov 2022. [Online]. Available: https://www.paulekman.com/universal-emotions/ xv, 28

[8] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980. xv, 27, 28

[9] R. Plutchik, "A Psychoevolutionary Theory of Emotions," *Social Science Information*, vol. 21, no. 4-5, pp. 529–553, 1982. xv, 28

[10] J. Murphy, U. Gretzel, and J. Pesonen, "Marketing robot services in hospitality and tourism: the role of anthropomorphism," *Journal of Travel & Tourism Marketing*, vol. 36, no. 7, pp. 784–795, Feb. 2019. 1

[11] P. Kenny, T. D. Parsons, J. Gratch, A. Leuski, and A. A. Rizzo, "Virtual patients for clinical therapist skills training," in *International Workshop on Intelligent Virtual Agents*. Springer, 2007, pp. 197–210. 1

[12] E. Matsas and G.-C. Vosniakos, "Design of a virtual reality training system for human–robot collaboration in manufacturing tasks," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 11, no. 2, pp. 139–153, 2017. 1

[13] B. Scassellati, J. Brawer, K. Tsui, S. Nasihati Gilani, M. Malzkuhn, B. Manini, A. Stone, G. Kartheiser, A. Merla, A. Shapiro *et al.*, "Teaching language to deaf infants with a robot and a virtual human," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13. 1

[14] S. Jeong, N. Hashimoto, and S. Makoto, "A novel interaction system with force feedback between real-and virtual human: an entertainment system:" virtual catch ball"," in *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, 2004, pp. 61–66. 1

[15] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 928–939. 1

[16] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, "Strategy and policy learning for non-task-oriented conversational systems," in *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, 2016, pp. 404–412. 1

[17] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and trends in human-computer interaction*, vol. 1, no. 3, pp. 203–275, 2007. 1

[18] K. Dautenhahn, "Socially intelligent robots: dimensions of human–robot interaction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 679–704, Feb. 2007. 1, 12, 13

[19] M. A. Goodrich and A. C. Schultz, "Human–Robot Interaction: A Survey," *Foundations and Trends in Human–Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2008. 11

[20] K. Dautenhahn, "Human-Robot Interaction," in *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* The Interaction Design Foundation, 2013. 12

[21] V. Lim, M. Rooksby, and E. S. Cross, "Social Robots on a Global Stage: Establishing a Role for Culture During Human–Robot Interaction," *International Journal of Social Robotics*, vol. 13, no. 6, pp. 1307–1333, 2021. 13

[22] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 143–166, 2003. 13

[23] C. L. Breazeal, *Designing Sociable Robots*. MIT press, 2002.

[24] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, "Understanding social robots," in *Proceedings of the International Conferences on Advances in Computer-Human Interactions*. IEEE, 2009, pp. 169–174. 13

[25] H. Yan, M. H. Ang, and A. N. Poo, "A Survey on Perception Methods for Human–Robot Interaction in Social Robots," *International Journal of Social Robotics*, vol. 6, no. 1, pp. 85–119, 2014. 13

[26] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, "Artificial cognition for social human–robot interaction: An implementation," *Artificial Intelligence*, vol. 247, pp. 45–69, 2017. 14

[27] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics*, vol. 3, no. 21, p. eaat5954, 2018. 15

[28] S. Rasouli, G. Gupta, E. Nilsen, and K. Dautenhahn, "Potential Applications of Social Robots in Robot-Assisted Interventions for Social Anxiety," *International Journal of Social Robotics*, 2022. 15

[29] R. de Kervenoael, R. Hasan, A. Schwob, and E. Goh, "Leveraging human-robot interaction in hospitality services: Incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots," *Tourism Management*, vol. 78, p. 104042, 2020. 15

[30] Z. Yumak and N. Magnenat-Thalmann, "Multi-party interaction with a virtual character and a human-like robot," in *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*, 2013, pp. 153–156. 15

[31] Z. Yumak, J. Ren, N. M. Thalmann, and J. Yuan, "Tracking and fusion for multiparty interaction with a virtual character and a social robot," in *SIGGRAPH Asia 2014 Autonomous Virtual Humans and Social Robot for Telepresence*, 2014, pp. 1–7. 15, 19

[32] D. Bohus and E. Horvitz, "Computational models for multiparty turn taking," Microsoft Research Technical Report MSR-TR 2010-115, Tech. Rep., 2010. 15, 18

[33] M. E. Foster, A. Gaschler, and M. Giuliani, "How can I help you': Comparing engagement classification strategies for a robot bartender," in *Proceedings of the ACM on International Conference on Multimodal Interaction*, ser. ICMI '13, 2013, pp. 255–262.

[34] Q. Xu, L. Li, and G. Wang, "Designing engagement-aware agents for multiparty conversations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2233–2242. 15, 29, 30

[35] D. Bohus and E. Horvitz, "Open-World Dialog: Challenges, Directions, and a Prototype," Microsoft, Tech. Rep. MSR-TR-2009-36, Apr. 2009. 15

[36] ——, "Learning to predict engagement with a spoken dialog system in open-world settings," in *Proceedings of the SIGDIAL Conference*, 2009, pp. 244–252. 16, 29, 31

[37] M. P. Michalowski, S. Sabanovic, and R. Simmons, "A spatial model of engagement for a social robot," in *Proceedings of the IEEE International Workshop on Advanced Motion Control*, 2006, pp. 762–767. 23, 29, 31, 42, 48

[38] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of the SIGDIAL Conference, The Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2009, pp. 225–234. 29, 42, 48

[39] D. Vaufreydaz, W. Johal, and C. Combe, "Starting engagement detection towards a companion robot using multimodal features," *Robotics and Autonomous Systems*, vol. 75, pp. 4–16, 2016. 30, 31, 42

[40] M. E. Foster, A. Gaschler, and M. Giuliani, "Automatically classifying user engagement for dynamic multi-party human-robot interaction," *International Journal of Social Robotics*, vol. 9, no. 5, pp. 659–674, 2017. 19, 30, 31, 42

[41] Y. Ozaki, T. Ishihara, N. Matsumura, T. Nunobiki, and T. Yamada, "Decision-making prediction for human-robot engagement between pedestrian and robot receptionist," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, 2018, pp. 208–215. 16, 29, 31, 42

[42] S.-S. Yun, "A gaze control of socially interactive robots in multiple-person interaction," *Robotica*, vol. 35, no. 11, pp. 2122–2138, 2017. 18

[43] R. Zhang, H. Lee, L. Polymenakos, and D. Radev, "Addressee and response selection in multi-party conversations with speaker interaction rnns," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 18

[44] N. M. Dowell, T. M. Nixon, and A. C. Graesser, "Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions," *Behavior research methods*, vol. 51, no. 3, pp. 1007–1041, 2019. 18

[45] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006. 18

[46] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4945–4949.

[47] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017. 18

[48] D. Traum and L.-P. Morency, "Integration of visual perception in dialogue understanding for virtual humans in multi-party interaction." in *International Workshop on Interacting with ECAs as Virtual Characters*, 2010, p. 70. 18

[49] D. Traum, "Issues in multiparty dialogues," in *Workshop on Agent Communication Languages*. Springer, 2003, pp. 201–211. 18

[50] D. Bohus and E. Horvitz, "Dialog in the open world: platform and applications," in *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009, pp. 31–38. 18

[51] R. Ishii, S. Kumano, and K. Otsuka, "Analyzing gaze behavior during turn-taking for estimating empathy skill level," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 365–373. 18

[52] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale rnns," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 186–190. 18

[53] D. Bohus and E. Horvitz, "Multiparty turn taking in situated dialog: Study, lessons, and directions," in *Proceedings of the SIGDIAL 2011 Conference*, 2011, pp. 98–109. 18

[54] ——, "Decisions about turns in multiparty conversation: from perception to action," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 153–160. 18

[55] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. P. Petrick, "Two people walk into a bar: Dynamic multi-party social interaction with a robot agent," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2012, pp. 3–10. 19

[56] Y. Kondo, K. Takemura, J. Takamatsu, and T. Ogasawara, "A gesture-centric android system for multi-party human-robot interaction," *Journal of Human-Robot Interaction*, vol. 2, no. 1, pp. 133–151, 2013. 19

[57] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004. 20

[58] J. K. Burgoon, L. K. Guerrero, and K. Floyd, *Nonverbal Communication*. Routledge, 2010. 20, 84

[59] R. Jones, *Communication in the real world: An introduction to communication studies*. The Saylor Foundation, 2013. 20

[60] E. T. Hall, *The Hidden Dimension*. Garden City, NY: Doubleday, 1966, vol. 609. 20, 23

[61] J. A. Hall and M. L. Knapp, *Nonverbal Communication*. Walter de Gruyter, 2013, vol. 2. 20, 84

[62] M. L. Patterson, "A Systems Model of Dyadic Nonverbal Interaction," *Journal of Nonverbal Behavior*, vol. 43, no. 2, pp. 111–132, 2019. 20, 84

[63] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal Communication in Human Interaction*. Cengage Learning, 2013. 20, 23

[64] M. Argyle, *The psychology of interpersonal behaviour*. Penguin UK, 1994. 20

[65] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009. 22

[66] A. Pentland, "Social Signal Processing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, 2007. 22

[67] V. Manusov and M. L. Patterson, *The Sage Handbook of Nonverbal Communication*. Sage Publications, Inc., 2006. 23

[68] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009. 23

[69] E. T. Hall, R. L. Birdwhistell, B. Bock *et al.*, "Proxemics," *Current Anthropology*, vol. 9, no. 2/3, pp. 83–108, 1968. 23, 48

[70] R. Mead, A. Atrash, and M. J. Matarić, "Proxemic feature recognition for interactive robots: automating metrics from the social sciences," in *Proceedings of the International Conference on Social Robotics*, 2011, pp. 52–61. 23, 48

[71] A. Kendon and A. Ferber, "A description of some human greetings," *Comparative Ecology and Behaviour of Primates*, vol. 591, no. 668, 1973. 23, 48

[72] S. R. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? cues to the direction of social attention," *Trends in cognitive sciences*, vol. 4, no. 2, pp. 50–59, 2000. 23, 49

[73] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, p. 92, 2020. 23

[74] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005. 23

[75] I. Poggi, "Mind, hands, face, and body: A sketch of a goal and belief view of multimodal communication," in *Body - Language - Communication*. De Gruyter Mouton, 2013, vol. 1, pp. 627–647. 24, 59

[76] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *Proceedings of the International Conference on Multimodal Interfaces*, 2009, pp. 119–126. 24

[77] P. Guhan, M. Agarwal, N. Awasthi, G. Reeves, D. Manocha, and A. Bera, "Abc-net: Semi-supervised multimodal GAN-based engagement detection using an affective, behavioral and cognitive model," *arXiv:2011.08690*, 2020. 32, 33, 34, 60

[78] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Transactions on Affective Computing*, 2021. 32, 33, 34, 60

[79] J. D. Finn and K. S. Zimmer, "Student engagement: What is it? why does it matter?" in *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds. Springer, 2012, vol. 840, pp. 97–131.

[80] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[81] J. Cohen-Mansfield, M. S. Marx, L. S. Freedman, H. Murad, N. G. Regier, K. Thein, and M. Dakheel-Ali, "The comprehensive process model of engagement," *The American Journal of Geriatric Psychiatry*, 2011.

[82] I. Archambault and V. Dupéré, "Joint trajectories of behavioral, affective, and cognitive engagement in elementary school," *The Journal of Educational Research*, vol. 19, no. 10, pp. 859–870, 2017.

[83] A. Ben-Eliyahu, D. Moore, R. Dorph, and C. D. Schunn, "Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts," *Contemporary Educational Psychology*, vol. 53, pp. 87–105, 2018. 60

[84] L. J. Corrigan, C. Peters, D. Küster, and G. Castellano, "Engagement perception and generation for social robots and virtual agents," in *Toward Robotic Socially Believable Behaving Systems*. Springer International Publishing, 2016, vol. 19, pp. 29–51. 29

[85] G. Perugia, M. Díaz-Boladeras, A. Català-Mallofré, E. I. Barakova, and M. Rauterberg, "ENGAGE-DEM: A model of engagement of people with dementia," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 926–943, 2022. 24, 25, 59

[86] C. Oertel, P. Jonell, D. Kontogiorgos, K. F. Mora, J.-M. Odobez, and J. Gustafson, "Towards an engagement-aware attentive artificial listener for multi-party interactions," *Frontiers in Robotics and AI*, vol. 8, 2021. 24

[87] J. Cohen-Mansfield, M. Dakheel-Ali, and M. S. Marx, "Engagement in persons with dementia: The concept and its measurement," *The American Journal of Geriatric Psychiatry*, vol. 17, no. 4, pp. 299–307, 2009. 25

[88] M. A. Trahan, J. Kuo, M. C. Carlson, and L. N. Gitlin, "A systematic review of strategies to foster activity engagement in persons with dementia," *Health Education & Behavior*, vol. 41, no. 1_suppl, pp. 70S–83S, 2014. 25

[89] C. Jones, B. Sung, and W. Moyle, "Engagement of a person with dementia scale: Establishing content validity and psychometric properties," *Journal of Advanced Nursing*, vol. 74, no. 9, pp. 2227–2240, 2018. 25, 73

[90] Y. Feng, G. Perugia, S. Yu, E. I. Barakova, J. Hu, and G. W. M. Rauterberg, "Context-enhanced human-robot interaction: Exploring the role of system interactivity and multimodal stimuli on the engagement of people with dementia," *International Journal of Social Robotics*, vol. 14, no. 3, pp. 807–826, 2021. 25

[91] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, 1992. 25, 86

[92] M. C. Ashton and K. Lee, "The HEXACO–60: A Short Measure of the Major Dimensions of Personality," *Journal of Personality Assessment*, 2009. 25, 86

[93] O. P. John and S. Srivastava, *The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives*. Guilford Press, 1999. 25

[94] P. T. Costa Jr. and R. R. McCrae, *The Revised NEO Personality Inventory (NEO-PI-R)*. Sage Publications, Inc, 2008, pp. 179–198. 26

[95] V. Benet-Martínez and O. P. John, "Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English," *Journal of Personality and Social Psychology*, 1998. 26

[96] B. Rammstedt and O. P. John, "Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German," *Journal of Research in Personality*, 2007. 26

[97] J. J. Gross and L. Feldman Barrett, "Emotion Generation and Emotion Regulation: One or Two Depends on Your Point of View," *Emotion Review*, vol. 3, no. 1, pp. 8–16, 2011. 27

[98] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Elsevier, 2013, vol. 11. 27

[99] A. Mehrabian, *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Cambridge: Oelgeschlager, Gunn & Hain, 1980. 27

[100] Y. Ozaki, T. Ishihara, N. Matsumura, and T. Nunobiki, "Can robot attract passersby without causing discomfort by user-centered reinforcement learning?" *arXiv preprint arXiv:1903.05881*, 2019. 29

[101] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A survey of autonomous human affect detection methods for social robots engaged in natural hri," *Journal of Intelligent & Robotic Systems*, vol. 82, no. 1, pp. 101–133, 2016. 29

[102] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "Towards engagement models that consider individual factors in hri: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task," *International Journal of Social Robotics*, vol. 9, no. 1, pp. 63–86, 2017.

[103] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 305–312. 29

[104] A. B. Youssef, C. Clavel, and S. Essid, "Early detection of user engagement breakdown in spontaneous human-humanoid interaction," *IEEE Transactions on Affective Computing*, 2019. 29

[105] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas, and P. Maragos, "A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task," *arXiv preprint arXiv:1812.00253*, 2018. 29

[106] H. Won Park, J. Busche, B. Schuller, C. Breazeal, R. W. Picard *et al.*, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 29

[107] Y. Matsuyama, I. Akiba, S. Fujie, and T. Kobayashi, "Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant," *Computer Speech & Language*, vol. 33, no. 1, pp. 1–24, 2015. 29

[108] K. Ito, Q. Kong, S. Horiguchi, T. Sumiyoshi, and K. Nagamatsu, "Anticipating the start of user interaction for service robot in the wild," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9687–9693. 32, 42

[109] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. 32

[110] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 32

[111] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions," *IEEE Access*, vol. 5, pp. 705–721, 2017. 32, 34, 35, 60

[112] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, 2019. 32, 36, 94, 102, 103

[113] A. Ben Youssef, C. Clavel, and S. Essid, "Early detection of user engagement breakdown in spontaneous human-humanoid interaction," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 776–787, 2019. 32, 33, 34, 60

[114] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, 2017. 33, 34, 49, 60

[115] N. Gao, W. Shao, M. S. Rahaman, and F. D. Salim, "N-Gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–26, 2020. 32, 33, 34, 60

[116] K. Saleh, K. Yu, and F. Chen, "Improving users engagement detection using end-to-end spatio-temporal convolutional neural networks," in *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 190–194. 32, 33, 34, 60, 76, 77

[117] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," *Frontiers in Robotics and AI*, vol. 7, p. 116, 2020. 33, 34, 60

[118] B. Zhu, X. Lan, X. Guo, K. E. Barner, and C. Boncelet, "Multi-rate Attention Based GRU Model for Engagement Prediction," in *Proceedings of the International Conference on Multimodal Interaction*, Oct. 2020, pp. 841–848. 33, 34, 60

[119] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, pp. 217–226. 33, 34, 60

[120] D. Anagnostopoulou, N. Efthymiou, C. Papailiou, and P. Maragos, "Engagement estimation during child robot interaction using deep convolutional networks focusing on asd children," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 3641–3647. 32, 33, 34, 60, 76, 77

[121] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012. 33

[122] L. Steinert, F. Putze, D. Küster, and T. Schultz, "Towards Engagement Recognition of People with Dementia in Care Settings," in *Proceedings of the International Conference on Multimodal Interaction*, Oct. 2020, pp. 558–565. 33, 34, 35, 60, 76, 77

[123] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 59–66. 35, 50, 67, 75, 96

[124] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, 2015, pp. 41.1–41.12. 35, 50

[125] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results," in *Computer Vision – ECCV Workshops*, 2016. 35

[126] F. Gürpinar, H. Kaya, and A. A. Salah, "Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2016. 35, 98

[127] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, "Apparent Personality Recognition from Uncertainty-Aware Facial Emotion Predictions using Conditional Latent Variable Models," in *Proceesings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2021. 35

[128] M. Romeo, D. H. García, T. Han, A. Cangelosi, and K. Jokinen, "Predicting Apparent Personality from Body Language: Benchmarking Deep Learning Architectures for Adaptive Social Human–Robot Interaction," *Advanced Robotics*, 2021. 36, 102, 103

[129] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features With 3D Convolutional Networks," in *Computer Vision – ECCV*, 2015. 36, 102

[130] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. v. Lier, "Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition," in *Computer Vision – ECCV*, 2016. 36, 98, 102

[131] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep Bimodal Regression for Apparent Personality Analysis," in *Computer Vision – ECCV*, 2016. 36, 98, 102

[132] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal First Impressions Recognition Using Temporally Ordered Deep Audio and Stochastic Visual Features," in *Computer Vision – ECCV*, 2016. 36, 102

[133] Z. Shao, S. Song, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Personality Recognition by Modelling Person-specific Cognitive Processes using Graph Representation," in *Proceedings of the ACM International Conference on Multimedia*, 2021. 36

[134] C. Palmero, J. Selva, S. Smeureanu, J. C. S. J. Junior, A. Clapes, A. Mosegui, Z. Zhang, D. Gallardo, G. Guilera, D. Leiva, and S. Escalera, "Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2021. 36, 82, 99

[135] H. Jin, Q. Song, and X. Hu, "Auto-Keras: An Efficient Neural Architecture Search System," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1946–1956. 36

[136] D. Curto, A. Clapés, J. Selva, S. Smeureanu, J. C. S. J. Junior, D. Gallardo-Pujol, G. Guilera, D. Leiva, T. B. Moeslund, S. Escalera, and C. Palmero, "Dyadformer: A Multi-Modal Transformer for Long-Range Modeling of Dyadic Interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021. 36

[137] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Joint Facial Action Unit Detection and Feature Fusion: A Multi-Conditional Learning Approach," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 572 715–335 742, Dec. 2016. 37

[138] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 37

[139] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time Convolutional Neural Networks for Emotion and Gender Classification," *arXiv:1710.07557*, 2017. 37

[140] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, May 2019. 37

[141] K. Schindler, L. Van Gool, and B. De Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008. 37

[142] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011. 37

[143] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion Recognition in Context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1667–1675. 37, 94

[144] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-Aware Emotion Recognition Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 143–10 152. 37, 94

[145] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 234–14 243. 37

[146] L. Zhang, S. Peng, and S. Winkler, "PersEmoN: A Deep Network for Joint Analysis of Apparent Personality, Emotion and Their Relationship," *IEEE Transactions on Affective Computing*, 2019. 38

[147] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125. 39

[148] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861*, 2017. 40

[149] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "MoViNets: Mobile Video Networks for Efficient Video Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 020–16 030. 40

[150] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019. 50

[151] Y. Kato, T. Kanda, and H. Ishiguro, "May I help you? - design of human-like polite approaching behavior-," in *Proceedings of the International Conference on Human-Robot Interaction*, 2015, pp. 35–42. 55

[152] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013. 55

[153] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "UE-HRI: A New Dataset for the Study of User Engagement in Spontaneous Human-robot Interactions," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2017, pp. 464–472. 55

[154] M. Ghafurian, J. Hoey, and K. Dautenhahn, "Social Robots for the Care of Persons with Dementia: A Systematic Review," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–31, 2021. 59

[155] "Ageing and health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, Oct. 2021. 59

[156] C. A. Liang, S. A. Munson, and J. A. Kientz, "Embracing four tensions in human-computer interaction research with marginalized people," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 28, no. 2, pp. 1–47, Apr. 2021. 60

[157] A. Abedi and S. Khan, "Affect-driven ordinal engagement measurement from videos," *arXiv:2106.10882*, 2021. 60

[158] G. Guo, R. Guo, and X. Li, "Facial expression recognition influenced by human aging," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 291–298, 2013. 60

[159] M. Fölster, U. Hess, and K. Werheid, "Facial age affects emotional expression decoding," *Frontiers in Psychology*, vol. 5, 2014. 60

[160] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555. 61, 64, 89

[161] A. Kamel, B. Sheng, P. Li, J. Kim, and D. D. Feng, "Hybrid refinement-correction heatmaps for human pose estimation," *IEEE Transactions on Multimedia*, vol. 23, pp. 1330–1342, Jan. 2021. 64

[162] S. Zeghoud, S. G. Ali, E. Ertugrul, A. Kamel, B. Sheng, P. Li, X. Chi, J. Kim, and L. Mao, "Real-time spatial normalization for dynamic gesture classification," *The Visual Computer*, vol. 38, no. 4, pp. 1345–1357, apr 2022. 64

[163] C.-F. R. Chen, R. Panda, K. Ramakrishnan, R. Feris, J. Cohn, A. Oliva, and Q. Fan, "Deep analysis of cnn-based spatio-temporal representations for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6165–6175. 64

[164] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," *arXiv:1705.06950*, 2017. 64, 96

[165] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," in *Computer Vision – ECCV*, 2022. 64

[166] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2961–2969. 67

[167] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203–5212. 67

[168] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6248–6257. 67, 75, 91

[169] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803. 68, 89

[170] E. Goffman, *Forms of talk*. University of Pennsylvania Press, 1981. 70

[171] H. H. Clark, *Using language*. Cambridge University Press, 1996. 70

[172] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008. 70

[173] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv:1710.10903*, 2017. 70

[174] N. Mishra, G. Tulsulkar, H. Li, N. M. Thalmann, L. H. Er, L. M. Ping, and C. S. Khoong, "Does elderly enjoy playing Bingo with a robot? A case study with the humanoid robot Nadine," in *Computer Graphics International Conference*, 2021, pp. 491–503. 73

[175] J. C. S. Jacques Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. van Gerven, R. van Lier, and S. Escalera, "First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis," *IEEE Transactions on Affective Computing*, 2022. 82

[176] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 82, 91

[177] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017. 82

[178] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 89

[179] M. Doyran, A. Schimmel, P. Baki, K. Ergin, B. Türkmen, A. A. Salah, S. C. J. Bakkes, H. Kaya, R. Poppe, and A. A. Salah, "MUMBAI: Multi-Person, Multimodal Board Game Affect and Interaction Analysis Dataset," *Journal on Multimodal User Interfaces*, 2021. 94, 104

[180] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," *arXiv:2107.08430*, 2021. 96

[181] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Transactions on Affective Computing*, 2017. 96

[182] F. T. Passini and W. T. Norman, "A Universal Conception of Personality Structure?" *Journal of Personality and Social Psychology*, vol. 4, no. 1, p. 44, 1966. 103

[183] D. Watson, "Strangers' Ratings of the Five Robust Personality Factors: Evidence of a Surprising Convergence with Self-Report," *Journal of Personality and Social Psychology*, vol. 57, no. 1, p. 120, 1989. 103

[184] A. Goldenberg, E. Halperin, M. van Zomeren, and J. J. Gross, "The Process Model of Group-Based Emotion: Integrating Intergroup Emotion and Emotion Regulation Perspectives," *Personality and Social Psychology Review*, vol. 20, no. 2, pp. 118–141, 2016. 110

[185] A. M. Nuxoll and J. E. Laird, "Extending Cognitive Architecture with Episodic Memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2007, pp. 1560–1564. 111

[186] R. Harel, Z. Yumak, and F. Dignum, "Towards a generic framework for multi-party dialogue with virtual humans," in *Proceedings of the 31st International Conference on Computer Animation and Social Agents*, 2018, pp. 1–6. 111

[187] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022. 111

[188] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," *Computer Speech & Language*, 2021. 111

[189] H. Ouchi and Y. Tsuboi, "Addressee and response selection for multi-party conversation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2133–2143. 111