Robot Behavior Generation and Human Behavior Understanding in Natural Human-Robot Interaction

Chuang Yu

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAE009

Thèse de doctorat

ENSTA IP PARIS

Robot Behavior Generation and Human Behavior Understanding in Natural Human-Robot Interaction

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale supérieure de techniques avancées

École doctorale n°626 de l'Institut Polytechnique de Paris (ED IPP)
Spécialité de doctorat : Informatique, données, IA

Thèse présentée et soutenue à Palaiseau, le 24/6/2021, par

CHUANG YU

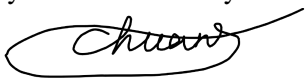
Composition du Jury :

Amel Bouzeghoub Professeure, Télécom SudParis IP Paris	Présidente du jury
Rachid Alami Directeur de Recherche CNRS, LAAS-Toulouse	Rapporteur
Emanuele Frontoni Professeur, Università Politecnica delle Marche, Italie	Rapporteur
Tony Belpaeme Professor, Ghent University, Belgique	Examinateur
Adriana Tapus Professeure, ENSTA-Paris IP Paris	Directrice de thèse

Declaration of Authorship

I, Chuang Yu, declare that this thesis titled, "Robot Behavior Generation and Human Behavior Understanding in Natural Human-Robot Interaction" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 20/07/2021

Acknowledgements

Time flies, four years have passed in a blink of an eye. I vaguely remember what I looked like when I first came to Paris. That day in 2017, I arrived in Paris with a huge luggage, which was as big as my dream. First of all, I would like to thank my supervisor Prof. Adriana Tapus. In the scientific research, she patiently taught me vast professional knowledge, how to start-up new research, and how to write a paper. She also inspired me significantly about cooperating with others, facing the pressure of scientific research, finding new ideas, and managing time well. In daily life, she also gave me and my fiancée a lot of care, allowing us to experience the warmth of home in a strange country. I have learned a lot from her and will never forget my experience as a PhD student in her lab. Thanks again to Prof. Adriana Tapus.

I would also like to acknowledge Prof. David Filliat, the director of the U2IS, ENSTA Paris, part of the Institut Polytechnique de Paris. It was his recommendation that allowed me to continue my Ph.D. dream here. At the same time, I enjoyed the experience at the Robotics and Autonomous Systems Laboratory, ENSTA Paris for its excellent scientific research environment, which greatly assisted my Ph.D. research. In addition, I would like to thank the colleagues and engineers in the laboratory for their collaboration in my projects, daily happy communication, and their enthusiastic help, which made me feel the warmth all the time.

In addition, I would also like to acknowledge my postgraduate tutor, Prof. Mei Shuai, who gave me massive bits of help and deepens my understanding and love of robots. At the same time, I am grateful to my undergraduate tutor, Prof. Huifen Dong, who lighted my dream of robotics research and allowed me to engage in robotics research in her robotics lab.

I would also like to thank the China Scholarship Council (CSC), whose funding made it possible for me to study in a famous robotics lab as a Ph.D. student. The CSC also funded my living expenses during my Ph.D. and provided free language training.

Furthermore, I would like to thank my scientific research partners and friends who have supported me a lot during my Ph.D. The days of communication with them have made me feel that life is no longer alone and scientific research is full of fun.

Finally, I would like to express great thanks to my fiancée Ya'nan for her companionship, support, and encouragements, which kept me striving to be better. At this moment, she is cooking a delicious dinner for me while I am writing the doctoral thesis. I will never forget this scene for the rest of my life. Moreover, I would like to express my sincere gratitude to my parents, sisters, and my extended family. Unforgettably, I still clearly remember how my mother accompanied me when I was doing homework under the lights at night. Unforgettably, I still clearly remember how my father picked me up from school. Unforgettably, I still clearly remember how happy I was with my sisters every day when I was a child.

Thank you all!

Contents

List of Figures

List of Tables

List of Abbreviations

HRI	**H**uman **R**obot **I**nteraction
GMM	**G**aussian **M**ixture **M**odel
HRC	**H**uman **R**obot **C**ollaboration
EEG	**E**lectro**e**ncephalo**g**ram
EMG	**E**lectro**m**yo**g**raphy
EOG	**E**lectro**o**culo**g**raphy
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
GRU	**G**ated **R**ecurrent **U**nit
LSTM	**L**ong **S**hort **T**erm **M**emory
TTS	**T**ext-**T**o-**S**peech
ASD	**A**utism **S**pectrum **D**isorder
DOF	**D**egree **o**f **F**reedom
AI	**A**rtificial **I**ntelligence
PD	**P**arkinson's **D**isease
RL	**R**einforcement **L**earning
GNN	**G**raph **N**eural **N**etwork
IMU	**I**nertial **M**easurement **U**nit
ANN	**A**rtificial **N**eural **N**etwork
HMM	**H**idden **M**arkov **M**odel
DTW	**D**ynamic **T**ime **W**raping
SVM	**S**upport **V**ector **M**achine
MLP	**M**ulti**l**ayer **P**erceptron
KNN	**K** **N**earest **N**eighbors
DOF	**D**egrees **O**f **F**reedom
GAN	**G**enerative **A**dversarial **N**etwork
S2FGAN	**S**peech to (**2**) **F**ace **GAN**
S2GGAN	**S**peech to (**2**) **G**esture **GAN**
IRL	**I**nteractive **R**obot **L**earning
FFT	**F**ast **F**ourier **T**ransform
DCT	**D**iscrete **C**osine **T**ransform
DTW	**D**ynamic **T**ime **W**arping
PCA-LDA	**P**rincipal **C**omponent **A**nalysis-**L**inear **D**iscriminant **A**nalysis
AU	**A**ction **U**nits
SVR	**S**upport **V**ector **R**egressors
LKJA	**L**eft **K**nee **J**oint **A**angle
RKJA	**R**ight **K**nee **J**oint **A**angle
LKJAV	**L**eft **K**nee **J**oint **A**ngle **V**elocity
RKJAV	**R**ight **K**nee **J**oint **A**ngle **V**elocity
LHJA	**L**eft **H**ip **J**oint **A**angle
RHJA	**R**ight **H**ip **J**oint **A**angle
LHJAV	**L**eft **H**ip **J**oint **A**ngle **V**elocity
RHJAV	**R**ight **H**ip **J**oint **A**ngle **V**elocity

NVIE	**N**atural **V**isible and **I**nfrared facial **E**xpression Database
PSD	**P**ower **S**pectral **D**ensity
RBF	**R**adial **B**asis **F**unction
GNN	**G**raph **N**eural **N**etwork
ST-GCN	**S**patial **T**emporal **G**raph **C**onvolutional **N**etwork
PCA	**P**rincipal **C**omponent **A**nalysis
BN	**B**ayesian **N**etwork
ROS	**R**obot **O**perating **S**ystem
PAM	**P**ick **A** **M**ood
SDK	**S**oftware **D**evelopment **K**it
DFT	**D**iscrete **F**ourier **T**ransform
ReLU	**Re**ctified **L**inear **U**nit
IRL	**I**teractive **R**obot **L**earning

Chapter 1

Introduction

1.1 Human-robot interaction in our Daily Life

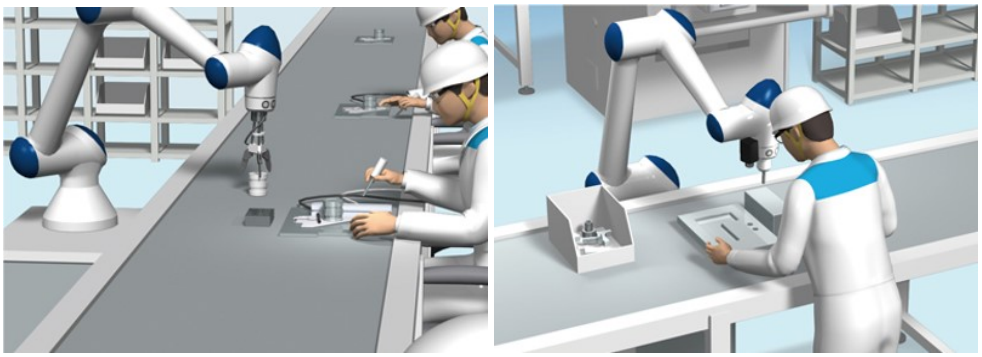
This thesis focuses on natural human-robot interaction (HRI), both from human behavior understanding and robot behavior generation side. Firstly, the role of Human-Robot Interaction (HRI) will be introduced. HRI is an interdisciplinary field, and brings together researchers and engineers from many areas, including computer science, robotics, philosophy, psychology, social sciences, etc. The interaction needs to consider and focus not only on the human but also on the robot. Before talking about human-robot interaction in daily life, the following two key research areas involved in HRI will be discussed:

- **How to perceive and understand human behavior?** This area focuses on the sensors and algorithms for human behavior perception. The sensors include cameras, audio recording devices, electroencephalogram (EEG) sensors, electromyography (EMG) sensors, pressure sensors, etc. Cameras are widely used in human-robot interaction. The RGB and RGB-D cameras can detect human body behaviour (e.g., gestures, postures, etc.), face action, and eye gaze information. This visual information can be used for human emotion understanding [1], human underlying intention detection [2], and health situation assessment [3] during human-robot interaction. Thermal cameras can extract the temperature-based emotional features from the human face infrared thermal imaging. These physiological features can be employed for example for human emotion recognition [4], and human lie detection [5] during human-robot interaction. EEG sensors can detect human brain's electrical activity signal. These signals can be decoded as the control information to control the robot in a physical human-robot interaction scene [6] or control an exoskeleton robot in a rehabilitation setting with HRI [7]. EMG signals correspond to the electrical activity in the human muscle. EMG signals are commonly used for human intention detection during human-robot collaboration [8], and the physical human-robot interaction for rehabilitation [9]. The pressure sensors can be used to determine user intention [10] and the human emotion detection during human-robot handshake [11]. Furthermore, there are different methods for human perception based on the data types. Convolutional Neural Network (CNN) [12] can process face and body for human emotion recognition [13] and human pose estimation [14]. Recurrent Neural Network (RNN), including Gated Recurrent Unit (GRU) [15], and Long Short-term Memory (LSTM) [16] can deal with human speech audio for emotion recognition [17], human speech text for chatbot [18], human body action for emotion recognition [19], and other types of time series in human-robot interaction scenes. Random Forest (RF) model, Support Vector Machine (SVM) model, and Hidden Markov Model (HMM) can also work on human speech or face emotion recognition [20] [21] [22]. Moreover, the transformer with the attention mechanism works well on the trimodal emotion recognition from speech text, audio, and vision [23].

- **How to conduct robot behavior?** Once the robot understands human behavior, it should have an appropriate response. During a human-robot conversation, the robot should learn how to speak with suitable speech text, speech audio, and natural non-verbal behaviors, such as synchronous robot face action, robot gesture, and robot head movement. There are two main methods for human-robot chatbot, namely the retrieval method and the generative method [18]. The seq2seq model [24] is mainly used for text sentence generation during robot response to human speech. The generated sentence should apply text-to-speech (TTS) technology to synthesize the human-like speech audio. Text-to-speech (TTS) generation research has recently made significant progress with some main methods [25], for example, Tacotron [26], Tacotron 2 [27], FastSpeech [28], and so on. Furthermore, style-controlled TTS is a new trend [29] [30]. This focus of research can enable a robot to have human-like speech audio with special characteristics during HRI, such as gender style, speed of speech, age, etc. Robots should also consider non-verbal behaviors during speech, such as robot face action, gesture, and head movement. Namely, a robot in a human-robot interaction scene should learn how to conduct the gesture relevant to the speech audio [31], or speech text [32] [33]. Some robots have their faces covered with skin, such as the ERICA robot [34] and the Zeno robot [35], and can actuate the face and eyes for human-like expressivity, which is vital in an effective and friendly human-robot interaction [36]. Furthermore, there is also the Furhat robot, possessing a face-swapping mask and an optical projection device behind the mask. The Furhat robot can change the face mask and customize the optical projection for limitless facial expressions with different personalities [37] [38]. These robot platforms can express rule-based limited number of facial expressions [39] and exhibit spontaneous face action aligned with robot speech audio and text during a friendly human-robot interaction. All in all, verbal and non-verbal robot behaviors greatly impact conducting an effective and human-like human-robot interaction [40]. The robot can imitate human behaviors [41] or learn by itself with reinforcement learning [42] in order to learn how to conduct appropriate robot behaviors or decision-making in a successful HRI.

A large number of advanced and intelligent robot platforms, including industrial robots for manufacturing and logistic robots for manipulation in a logistic network, have been used effectively for many years. These robots improved the quality of workplace and led to major productivity gains and profitability for many factories and industrial companies. They primarily work in a structured and constrained environment by having repeated movements and behaviours all the time. However, there are still many unstructured environments that need a robot to interact with the environment and humans in a safe and trustworthy manner. For example, assistive robots that help the elder living independently [43], exoskeleton robots that improve the rehabilitation programme of a disabled at home [44], and rescue robots that are used to substitute humans in a search scene or in dangerous environments [45]. It is challenging to put the robot in an unstructured setting with human interaction because humans always have a high uncertainty level at different times and places. Hence, there is still a lot of work to do to improve the level of maturity of human-robot interaction. Nowadays, human-robot interaction (HRI) research has attracted more and more new researchers worldwide with the development of the artificial intelligence algorithm and robot hardware platforms. Many human-robot interaction results are promising to be applied in the real-world setting.

Industrial robots have been used in many setups for many years, such as car and cellphone assembling with robot arms and the goods manipulation with logistic robots. With the ”Industrial 4.0” age coming, more collaborative robots working with humans have been used in industrial settings, namely cobots in industrial settings [46] [47]. They are more and more integrated because they can complete more complicated operations, and human-robot



(a) Collaborative quality inspection. (b) Collaborative product assembling.

Figure 1.1: Applications of the collaborative industrial robots with human-robot interaction [49].

collaboration solutions can make full use of humans' high-level cognitive capabilities and robots' excellent speed, repeatability, and accuracy [48]. Yaskawa Electric's Collaborative Robot MOTOAN-HC Series can be used for many safe and friendly human-robot collaboration/interaction tasks without a safety fence, such as collaborative quality inspection and collaborative product assembling, as shown in Figure 1.1 [49]. Humans are not good for producing accurate repetitive behaviours for long periods of time. Furthermore, humans can change their behaviour mode in a industrial human-robot interaction environment. In this context, the robot should predict the human action in order to conduct a safe HRI task. Maria et al. [50] built-up a robot learning framework with multiple demonstrations, in which the robot can learn the action series of an assembly task without pre-programming the robot trajectory planning. Besides, they also came up with a new adaptive algorithm robot learning based on the Gaussian Mixture Model (GMM). It can enable the robot to deal with the changes during the real-time HRI or Human-Robot Collaboration (HRC) scene, such as the position change of the object to be manipulated and human moving. Furthermore, it is promising to be used in a continuous HRI with an industrial collaborative robot. Multimodal human behavior signals should be considered to conduct a friendly and safe human-robot interaction in an industrial process [48]. Furthermore, in recent researches, human brain electroencephalogram (EEG) signal [51], electromyography (EMG) signal [52], gesture [53], and speech audio [54] have widely been used to interact with an industrial robot.

With the development of deep learning, machine vision, and sensor technology, HRI plays a more and more crucial role in the medical or healthcare context [55]. When working with vulnerable populations, the HRI becomes even more challenging. In [56], the AI-robot company and Beihang University have created an exoskeleton robot, which can help disabled people with lower-limb gait rehabilitation. The robotic system also includes a patient motion assessment subsystem with vision-based gait information from a depth camera and the foot plantar pressure signal from the thin film pressure sensors, which can assess the rehabilitation status of patients using the exoskeleton robot [57] [58]. The robotics researchers from Beihang University and Aalborg University build-up a lower-limb exoskeleton rehabilitation robotic system, which can be used to rehabilitate stroke patients, paralyzed patients, and spinal cord injury patients by using physical human-robot interaction. Moreover, the exoskeleton robot system also considers the EEG signal from the patient brain and EMG from the patient's lower-limb muscles to make the patient feel the exoskeleton robot as a part of the body, which can result in a friendly and natural physical human-robot interaction and can fasten the rehabilitation process [56]. In the paper [59], the authors employed multimodal human physiological information such as Electrooculography (EOG) and Electromyography

(EMG) to guide the robot to do reach-to-grasp task with real-time human-robot interaction. The EOG signal was decoded to complete the robot reaching movement towards the object manipulated, and the EMG signal was decoded to perform the robot grasping object task. The project is promising to assist disabled people in daily life with reliable classification accuracy.

Furthermore, nowadays, the number of people aged 60 years and older is growing and there is a shortage of primary care services. The numerous older adults live alone and experience cognitive and physical decline. The advancement in the human-robot interaction area also makes it possible for the assistive robot with HRI to help this vulnerable population in their own home environments with daily health monitoring, and social care [60]. The Europe Horizon 2020 project - Enabling Robot and assisted living environment for Independent Care and Health Monitoring of the Elderly (ENRICHME) focused on the physiological behavior monitoring with non-invasive sensors and adaptive HRI for older people assisting at home [61] [62] [63]. The human-robot interaction scene in the project ENRICHME is as shown in Figure 1.2. The robot TIAGo used in the project ENRICHME was equipped with a thermal camera and a Kinect depth sensor to detect the behaviors and physiological parameters of the users [64]. The project developed nine cognitive games for the robot platform, including Stroop game, Speed game, Memory game, Puzzle.

Figure 1.2: Human-robot interaction with the elderly in the project ENRICHME. The robot TIAGo monitors the health situation of the elderly and interact with the old people [64].

Humans can make full use of social intelligence in an efficient and successful bi-directional communication scenario. A robot should do the same in HRI [65]. Many robots are also used in domestic settings or public spaces [66] [67]. A social robot with effective interaction capabilities can potentially contribute to diagnosing and treating some particular diseases, such as autism spectrum disorder [68] and Parkinson's disease [69] [70]. Children with Autism Spectrum Disorder (ASD) may communicate with others with restricted behaviors and limited understanding of others' feelings in everyday social interaction. Many social robots have been developed for children with ASD [71], [72], [73], for example the robots NAO, Kaspar, Keepon, FACE, Probo, Charlie, Zeno, and CuDDler, as shown in Figure 1.3. The robot Zeno possesses an expressive robot face covered with skin and conversational artificial intelligence (AI). Marinoiu et al. applied the Zeno robot in the human-robot interaction context with human pose perception and emotion recognition of ASD children [74]. The robot NAO was used to play games with autistic children for social engagement and ASD treatment [75]. An example of interaction between the Kaspar robot and a child with ASD is shown in Fig. **??**.

Figure 1.3: The social robots with human-robot interaction for ASD diagnosis and intervention. **a:** robot NAO, **b** robot KASPAR, **c** robot Keepon, **d** robot FACE **e** PROBO robot, **f** robot CHARLIE, **g** Zeno robot, **h** CuDDler robot.

People suffering of Parkinson's disease (PD) suffer from social communication problems with limited expression of emotional face, gesture, and speech. The authors in [76] described a social robotic system designed for PD persons. The robot system can detect their emotions by using the speech modality and express emotional behaviors with the other non-verbal modality, namely gesture, which can considerably improve the PD persons' life quality during social communication.

Furthermore, with the development of internet technology and AI, the education area has been revolutionized by distance education. For example, students and researchers learn new knowledge or join workshops through video conferences and virtual meetings during Covid-19 pandemic. Compared to the virtual education method, the telepresence robots for education are being studied by various researchers [77], [78], [79]. The authors in the paper [80] explored the usefulness of the telepresence robot with human-robot interaction for language education. In [81], the authors tried to use the telepresence robot to help the K-12 (12th grade) students who cannot go to school for education because of various reasons. Robots for education can help children to learn new knowledge but can also be an useful aid to the teachers for professional development [82]. The European project EMOTE designed and developed a tutor robot possessing empathy, which plays a vital role in establishing relationships [83]. Such an example is shown in Fig. 1.4.

Figure 1.4: The HRI scenes with empathic education robot tutor [83].

1.2 How to achieve a natural human-robot interaction?

1.2.1 Overview of natural human-robot interaction

Naturalness is a significant index for social or service robots in a successful and trustworthy human-robot interaction setting. Naturalness in human-robot interaction as a multifaced concept faces multiple challenges when the current social or service robots tend to work well on a single task and have a limited level of multimodal perception and interaction ability. Hence, there are many works on natural HRI that should be explored. However, what is a natural human-robot interaction? It is not easy to give a precise definition. From the paper [84], it is clear that a natural human-robot interaction should consider natural controllability and natural behavior and also require a naturally controlled robot that can perform natural and expected behaviors to respond to human interaction.

Humans naturally communicate with each other through verbal behavior with speech and non-verbal behaviors, including facial expression, body gesture, gaze, social distance, and so on. A person can understand the interlocutor behaviors easily and also can conduct appropriate behaviors in different scenes. Namely, natural human-human communication needs bi-directional contribution. Accordingly, natural-human-robot interaction should be a two-way process where a robot should understand human commands and respond with natural robot behaviors. For example, an assistive robot designed for elderly persons should recognize human needs through human behavior understanding algorithms and then propose several care services. The authors in [84] define this kind of robot as a naturally controllable robot. The naturally controllable robot can be controlled through the perception of natural communication modalities, such as human gesture interface [85] [86], verbal speech interface [87], multimodal interface [88], etc. In addition, a robot also should learn to perform natural behaviors through verbal and non-verbal modalities in natural human-robot interaction, such as robot speech [89], gesture [90], face action [39], gaze control [91], and social distance (proxemics) [92].

The naturally controlled robots have been studied by many researchers in the past few years. Multiple natural communication modalities can be used for a naturally controlled robotic system, such as human speech, gesture, and gaze.

Research on developing natural robot behaviors gains more interest every year. Human speech can convey emotion, intention, and verbal content and it is essential for natural interpersonal communication. Non-verbal behaviors are neglected commonly during human-human communication. However, non-verbal behavior shares 60 % to 70 % contribution of human communication [93]. Hence, natural-behavior robots should be endowed with natural verbal and non-verbal behaviors. The detailed content and role of non-verbal behavior in natural human-robot interaction will be described in 1.2.3.

Emotional speech synthesis as an active research area has attracted massive attention from researchers around the world. The extensive development of text-to-speech and natural language processing technology has made it possible to generate realistic and emotional speech in the past few years [94] [95], which facilitates the social robot to express emotional states in speech. The authors in [96] presented an emotional speech synthesis system to empower the healthcare robot with empathy during human-robot interaction. The emotional speech synthesis system is based on the Random Forests and Adaboost, where emotion tag and text are inputs and the emotional speech for a social robot is the output. The HRI experimental results with emotional speech show that the successful empathetic robot speech requires not only the primary emotions but also the secondary emotions. The paper also indicates that the empathetic robot with emotional speech makes a good contribution to the acceptable human-robot interaction.

The paper [97] explored the personality of the synthesized speech based on the Big-Five personality questionnaire. Similarly, the robot behaviors also should own the personality feature to some degree. In the past research, Lee et al. [98] had certified that the participants preferred communicating with the robot with a complementary personality to their personalities than the robot with a similar personality with themselves. Aylett et al. [99] conducted a pilot study to explore the personality of generated robot speech. The generated robot speech with personality has significant potential in natural human-robot interaction.

The past research concluded that gender bias happened during human-computer interaction with verbal speech [100]. The paper [101] explored how the gender feature of robot speech influences the persuasiveness in the human-robot interaction with female and male users. NAO robot was endowed with gender features with the help of male and female speech utterances during the experiments. The research results show that a male speech robot is more persuasive than a female speech robot. The experiment results also indicate that the female users believe the speech robot with more persuasiveness than male users during human-robot interaction. All in all, gender property plays an essential role in natural human-robot interaction.

Moreover, interpersonal communication usually applies multiple natural modalities which convey richer information compared to unimodal one. For example, the human video chat with speech, face, gaze, and gesture makes a more natural and reliable communication than the audio chat. Similarly, multimodal communication also makes a difference in natural human-robot interaction [84]. It contains two parts, namely the multimodal human behavior perception and multimodal robot behavior expression in HRI.

Humans always perceive the surroundings with multimodal information because multimodal perception can make a better performance. So does the robot. The paper [102] described a multimodal human action recognition system in the context of human-robot interaction between an assistive robot and an elderly user with a mobility disability. The action recognition system is for the human command perception used in robot control. The multimodal information used in the action/command system included visual gesture images extracted from the Kinect camera and acoustic data obtained from the microphones. The results indicated that the multimodal command recognition system got better than the unimodal one, which possibly leads to a more successful HRI.

Humans tend to perform multimodal behaviors with speech, gesture, and facial expression in natural interpersonal communication. In the same way, multimodal robot behaviors can make an excellent contribution to a successful human-robot interaction. The paper [103] compared the HRI effects between the multimodal robot behaviors with speech and spontaneous gesture and unimodal robot behavior only with the speech in a situational context with the Honda humanoid robot. The experimental results show that the users prefer giving a more positive assessment to the HRI with multimodal robot behaviors than HRI with unimodal one, which can inspire the social robot behavior design for a natural HRI in the future.

1.2.2 The role of emotion in naturally social interactions (H-H and go to H-R)

Emotion can be a personal feeling when you live alone and think about some things. And emotion is also interpersonal when you interact with others with emotional contagion. Human emotion is significant in interpersonal communication because it can bone the social relationships between interactors. If your friends share their experiences with emotions with you, your excellent perception of the counterpart's emotion and appropriate emotional response will contribute to a close interpersonal relationship. Hence, a natural and successful human-human interaction needs the excellent ability of emotional perception and emotional behavior expression.

During human-human communication, a person can detect counterpart emotion from unimodal or multimodal human behaviors, including human facial expressions, semantic speech, non-verbal vocalization, and bodily action. A human also can conduct an expressive behavior to show his or her emotional internal state for natural human-robot interaction. For example, the person can speak with a happy or a sad voice spontaneously with emotional facial expressions and gestures. In addition, humans can do well in emotion adaptation, where people deal with giving more appropriate and adaptive responses during the interaction. All in all, a natural emotion-based human-human interaction contains three parts as follows.

1. how to perceive human's emotion well.
2. how to select the appropriate emotional behaviors to respond to the perceived emotion.
3. how to conduct the selected emotional behaviors.

Similarly, emotion plays a significant role in human-robot interaction. Emotion can help robots receive more attention from the human user and augment human engagement during human-robot interaction [104]. Furthermore, emotion contributes significantly to the continuous social presence in long-term human-robot interaction [105]. Natural affective human-robot interaction also includes successful human emotion recognition by a robot, a suitable reasoning and action selection mechanism, an appropriate robot expression platform and an effective robot behavior synthesis model [106] as shown in Figure 1.5. The emotion elicitation part in the figure is commonly used to trigger desired emotions of users in the laboratory settings. The popular emotion elicitation methods contain the image-based method, music-based method, film-based method, etc. [107]. However, in the realistic and in-the-wild human-robot interaction, it is no need to do emotion elicitation. Emotional behavior generation is divided into three parts in that paper [107]. The emotion expression part is related to different possible emotion expression modalities on a social robot. For example, the robot with an arm and skinned face can conduct affective facial expressions and emotional gestures during HRI. The robot with a vocalization system can speak with an emotional speech. The emotion adaptation part is about how to build a suitable behavior mechanism for response to the detected emotion from the human interactor. The emotion synthesis part is relative to generating or synthesizing the emotional robot behaviors, which are advised in the emotion adaptation part.

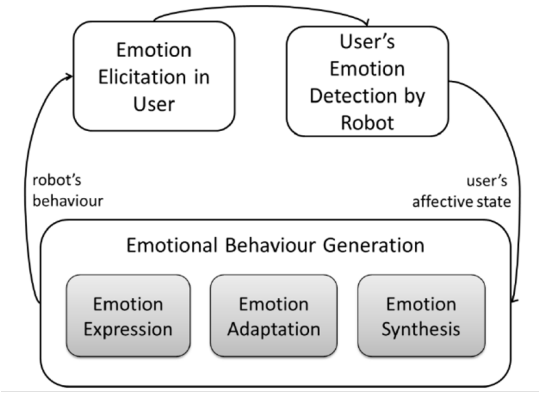


Figure 1.5: Affective Loop of Emotional Robots. Extracted from the paper [106]

There are many modalities that can be used to build up the human emotion recognition system for HRI, such as speech with text or audio, human bodily action, facial action, tactile

sense, physiological signals, and multimodal information. These human emotion recognition systems can allow the social robot to interpret human inner emotional state, which facilitates a friendly and natural human-robot interaction.

Emotion perception through speech audio is more effective with less computation than a vision-based model [108]. Therefore, it is a primary emotion recognition method used in human-robot interaction. The paper [109] explored the speech-audio-based emotion recognition system to classify basic emotions for human-robot interaction. The paper used feature selection processing to reduce the computational complexity. The authors also tried six classifiers, including Support Vector Machine (SVM), k Nearest Neighbors (kNN), decision tree, Bayesian Belief Networks, Naive Bayes classifier, and Multilayer Perceptron (MLP). The results show that SVM and Bayesian Belief Networks get better offline testing accuracy. Because the two models are also lightweight, they are possibly the candidate emotion recognition model in real-time human-robot interaction. Speech text is also a vital tool to sense human emotion. Yoon et al. [110] came up with a multimodal speech emotion recognition model with audio and text. The model based on dual recurrent neural networks (RNNs) can classify basic emotions (angry, happy, sad, and neutral) and obtained better emotion recognition performance, which inspires the future human emotion model in a human-robot interaction scene.

The face is an important modality for expressing human emotions. Face-based human emotion perception is helpful in natural HRI. Researchers have used multiple types of face data for human emotion recognition, for example, RGB face images [111], depth images [112], and thermal images [113]. The paper [111] explored facial emotion recognition with a NAO robot in the children-robot interaction. The paper compared the CNN-based emotion classification model and AFFDEX SDK [114] as an open facial emotion recognition platform. The CNN model showed a better emotion classification ability during the human-robot interaction. It is a potential emotion model used for a real-time and natural HRI in the future.

Emotion based robot behavior also makes a valuable contribution to a successful and natural human-robot interaction. Past research has shown that a robot endowed with an emotional behavior can be perceived as empathetic [96] and trustworthy [115], which are important features for natural HRI. These behaviors contain emotional speeches, facial expressions, and emotional body action. The paper [116] has shown that a robot could convey emotion with robot face action, robot head or locomotion movements, and robot gestures as well.

Chella et al. [117] built-up an emotional storyteller robot who can tell the story with the prosody-based emotional speech. The paper [118] explored the effect of emotional robot behaviors in children-robot interaction. During the experiments, the robot can express emotion with its speech, gesture, posture, eye colors, and bodily poses. The questionnaire results from children in HRI show that children enjoy interacting with a robot with adaptive emotion and gesture more than the robot without adaptive emotion and gesture. The paper [119] introduced the robot EmIR for the elderly assistance. EmiR can detect the users' emotions. It can also express empathy responding to the detected emotion with the help of visual cues, including the cheek color controlled by LED emitters and the facial expression, which leads to an acceptable human-robot interaction.

1.2.3 Non-verbal behaviour for natural human-robot interaction)

Nonverbal behaviors are essential in interpersonal communication. The verbal behavior is often conducted with the nonverbal behaviors in natural human-human interaction. Nonverbal behaviors can help to augment the communication engagement and attract continuous counterpart attention in long-term interaction. They also make the interaction more expressive. For instance, When a speaker gives a speech in public, the speech with rich gestures and

facial expressions will be more attractive and expressive than one without (or with limited) nonverbal behaviors. Nonverbal behaviors can lead to more versatile communication than verbal behavior to some degree. For example, persons who cannot speak the same language can share information through gestures or hand motions. Based on the reference [120], the categories of non-verbal behaviors are as shown below:

(1) face expression
(2) gaze
(3) gestures and other bodily movements
(4) posture
(5) bodily contact
(6) spatial behavior
(7) clothes and other kinds of appearance
(8) non-verbal vocalization
(9) smell

Nonverbal robot behaviors are essential and can facilitate a natural human-robot interaction [40]. The paper [121] shows that nonverbal robot behavior can promote task performance in a difficult collaborative task. The author designed the memorization task with different difficulty levels. In the collaborative task, the users worked with the robot with or without the nonverbal behaviors. The experimental results demonstrate that nonverbal cannot contribute to user performance in the lower-level-difficulty memorization task but can improve the user performance in the high-level-difficulty task.

Like human communication, the humanoid robot arms can conduct gestures and poses, and the skinned face robot can express facial action for a natural HRI. However, the robot also can own other non-verbal behaviors that humans cannot have, for example, the robot eye color controlled by LED light [118]. The nonverbal robot behaviors include robot facial action, robot gestures, robot gaze, robot contact, robot smell, non-verbal robot vocalization, robot appearance, robot proxemic behavior, or other possible ones.

Robot facial behaviors have two types: facial expression for emotion demonstration and co-speech facial action. They are both critical in natural HRI. The paper [122] explored the robot facial expression design of a companion robot for children, which can lead to good human-robot interaction. Robot nonverbal vocalization includes laughter, cry, sign, and so on, which can convey emotion easily and intuitively to users during a natural HRI. The paper [123] presented a nonverbal-vocalization-driven robot facial action generation system for the multimodal nonverbal HRI. With the help of the synthesis system, the robot ERICA can conduct the non-verbal vocalization (laughter and surprise) and the aligned face action spontaneously, which are vital for the naturalness of human-robot interaction.

The gesture contains co-speech gestures and speech-independent gestures, which play a vital role in human social behaviors and natural human-robot interaction. The paper [33] explored robot gesture generation with speech and showed that the interactors preferred the interaction with gesture and speech instead of only speech during human-robot interaction. Ondras et al. [124] used deep learning (LSTM and MLP) to generate the co-speech robot upper-body motions, including head, arm and torso movements. The generated robot gesture and speech were applied to the robot Pepper robot for experiments. Those generated co-speech robot gestures are possible to be used in future multimodal natural HRI.

The robot gaze also plays a central role in guiding user's attention and regulating turn-taking during interpersonal communication. The robot gaze behaviors include the robot gaze with one person and the robot gaze in group communication. The paper [91] explored the responsive robot gaze during human-robot interaction. The robot eye attention relies on what the user looks at in a responsive robot gaze scene. The HRI evaluation experiments showed that the responsive robot gaze experienced a stronger feeling of being looked at than the non-responsive robot gaze. The paper [125] explored the robot gaze in group communication.

The HRI experimental results showed that the gaze behaviors under consideration influenced the users' understanding of the robot action and that robot action influenced the human perception of the robot gaze, which inspires the natural HRI with the robot gaze.

Robot contact or robot haptic communication in HRI has many types, for example, human-robot hug and human-robot handshaking. It refers to how a robot interacts with the human through touching. It is a vital interaction modality in natural HRI. The paper [126] developed a human-robot handshaking system. And using the system, the authors conducted the HRI experiments to explore tactile-based emotion detection.

Robot smell was rarely studied in the past few years, even though human olfactory communication is vital during interpersonal communication [127]. The paper [127] endowed the social robot with an olfactory display system that can emit the smell to decorate the robot movement. The paper completed the synchronization between the controllable smell presentation and the robot movements in human-robot interaction. The research is inspiring for future natural HRI with smell representation.

Robot appearance also can make a difference in natural human-robot interaction. For example, the human-like robot ERICA was designed with high-level anthropomorphism [128]. ERICA has a skinned face, skinned hand, and skinned body, facilitating more natural human-like robot behaviors. The Furhat robot [37] has a back-projected plastic face and can replace the face quickly, but its robot face suffers from the influence of light compared with the ERICA robot. The ERICA robot and Furhat robot are as shown in Figure 1.6. The paper [129] conducted a survey and identified 157 robots with rendered faces and coded them in terms of 76 properties to explore the relation between people's perceptions of rendered robot faces and different facial features. The results showed that the presence or absence of specific features influences the human perception of the robot face, which can guide the robot face design for natural human-robot interaction in the future.

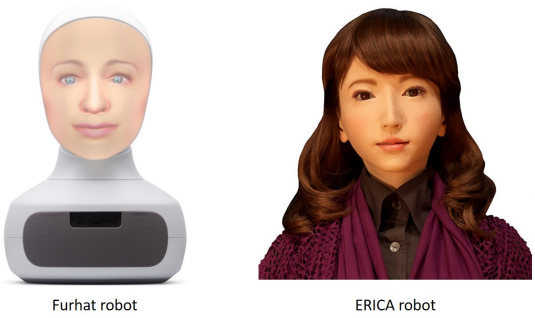


Figure 1.6: The robot Furhat (left) and the robot ERICA (right).

Robot proxemic behavior is about keeping a social distance during human-robot interaction, which is essential in social human-robot interaction [130]. The paper [131] came up with a real-time socially-aware robot navigation system, namely *SocioSense*, among pedestrians. The system takes the video with streaming pedestrians as input and outputs the proxemic distances and predicted trajectories for the socially aware robot navigation, which will lead to a safe and successful human-robot interaction.

All in all, the nonverbal behaviors contribute significantly to a successful human-robot interaction. A natural HRI should combine multimodal behaviors with verbal behaviors and non-verbal behaviors. Aly et al. [132] built up a multimodal behaviors synthesis system with speech, facial expression, and gestures for the ALICE robot. The experimental results validate the effectiveness of the generated multimodal robot behaviors in human-robot interaction.

1.3 Current Challenges in Social Robotics and Research Questions

A social robot is an intelligent robot that interacts with humans through social interaction [133] and can be a particular member of our society in the future [134]. Social robotics as an interdisciplinary is a new branch of robotics and has caught the emerging attention of researchers from various disciplines around the world [135]. Instead of an inanimate machine like the industrial robot in a traditional factory, the social robot can communicate and interact with a human with social cues and a social robot may possess a certain degree of humanness in the future [134]. A social robot should know how to perceive humans' social behaviors and how to respond to humans with social behavior during social human-robot interaction. For example, a social robot should recognize the emotion of the human counterpart well in an emotional interaction scene. It should also learn how to conduct emotional interaction with verbal and nonverbal behaviors.

Social robotics is popular now around the world, which attracts the massive attention of researchers. It can provide social assistive services to the particular group, such as children with a social disability and the elderly with cognitive disorder. Many robot platforms are capable of providing social service, for example, the robot NAO [136], the robot PARO [137], the robot Kaspar [138], and the robot Furhat [37]. The paper [139] used a socially expressive robot Pleo to assist children with autism spectrum disorders to develop social behaviors. And some other authors [140] explored the effectiveness of social robots including the robot NAO and the robot PARO in the therapy of advanced dementia.

However, many challenges of social robotics still lie ahead. Many social robots still conduct handcrafted behaviors during socially human-robot interaction experiments.Inspired by the reference [133], in this thesis we will describe the social robot challenges from the following views based on this thesis focus.

(1) Embodiment: challenging to build embodied robot for satisfactory performance of behavioral expression.

(2) Personality: challenging to inject personality into the robot's behavior with the learning method.

(3) Empathy: challenging to work well in in-the-wild or online emotion recognition; challenging to respond to the human emotion with appropriate robot behavior autonomously.

(4) Engagement: challenging to attract the user's attention for long-term HRI.

(5) Adaptation: challenging to adapt behaviors in long-term social human-robot interaction.

Facing these challenges, my thesis explores four research questions:

(1) how to accurately perceive human emotion recognition during human-robot interaction with multimodal signals, including thermal facial images and 3D gait data?

(2) how to build a long-life human emotion recognition model for reliable online perception performance?

(3) how to model an one-to-many generative model for natural co-speech robot gesture synthesis task?

(4) how to generate the robot face action align with speech audio for natural human-robot interaction?

1.4 Thesis contributions and architecture

This thesis aims to facilitate a natural human-robot interaction based on human behavior perception and robot behavior generation. We explored multimodal emotion recognition with the thermal facial features and 3D gait features during human-robot interaction in terms

of human behavior perception. For robot behavior generation, we completed an one-to-many speech-driven robot gesture generation. We also conducted the speaking robot face action generation for the social robot with the skin-covered face. The generated robot behaviors, including the talking face actions and co-speech gestures, were also applied to the robot's face (Zeno) and the humanoid robot (Pepper). Hence, the thesis contains four sub-projects: (1) one-to-many speaking robot gesture generation, (2) co-speech robot face action generation, (3) multimodal emotion recognition, and (4) interactive robot learning for long-life emotion perception. The related contributions are as follows, respectively:

(1) An audio-visual database based on the public TED videos. The database contains speech audio data and the aligned 3D human gestures obtained from 2D images of TED videos. A proposed one-to-many temporal Generative Adversarial Network (GAN) framework of the cross-modal generation is proposed to align the human-like gestures and the speech. The one-to-many means that the GAN model fed with one speech input will output multiple gestural time series. The generator model of GAN also introduces a new autoencoder architecture to solve the problem of adding the noise for the generative model with time series source as the input because the temporal noise sequence is hard to generate for temporal GAN. Finally, the generated co-speech gestures are mapped to the real robot Pepper, which certifies the generated gesture's effectiveness in reality.

(2) A temporal GAN architecture with L1 reconstruction loss was proposed to synthesize 3D co-speech face action effectively. A face action retargeting task was performed from human 3D face action to the robot face actuators. The generated speaking robot face actions are also employed on the real robot Zeno. Finally, user experiments have been done to assess the effectiveness of the generated speaking robot face action.

(3) The human-robot interaction experiments with human multimodal behaviors have been performed to build up a multimodal database with 3D gait data and thermal facial images. And the feature extraction tasks are done on the multimodal data. Then, CNN, HMM, SVM, RF, and a proposed hybrid model with gait and thermal facial features are explored in our database. The results certify the proposed hybrid model's effectiveness in multimodal emotion recognition.

(4) A new hybrid model based on Random Forest (RF) and confusion matrices of two unimodal RF models was developed. In addition, the interactive robot learning (IRL) architecture with the human in the loop is proposed, where the human verbal feedback is utilized in the long-term learning scene to improve the performance of real-time emotion recognition. The experiment results show the IRL method is efficient in a long-life multimodal emotion recognition scene.

The rest thesis is organized as follows:

Chapter 2 describes the one-to-many co-speech robot gesture generation with a temporal generative model.

Chapter 3 elaborates the speaking robot face action generation with a generative model.

Chapter 4 details the multimodal human emotion recognition with a thermal camera and RGB-D camera in human-robot interaction.

Chapter 5 presents the interactive robot learning for a better online multimodal emotion recognition for natural human-robot interaction.

Chapter **??** concludes my thesis contribution and possible works that can be proposed in the future.

Chapter 2

Robot gesture synthesis based on speech

2.1 Overview

The human gestures occur spontaneously and usually they are aligned with speech, which leads to a natural and expressive interaction. Speech-driven gesture generation is important in order to enable a social robot to exhibit social cues and conduct a successful human-robot interaction. In this chapter, the generation process involves mapping acoustic speech representation to the corresponding gestures for a humanoid robot. Our work proposes a new GAN (Generative Adversarial Network) architecture for speech to gesture generation. Instead of the fixed mapping from one speech to one gesture pattern, our end-to-end GAN structure can generate multiple mapped gestures patterns from one speech (with multiple noises) just like humans do. The generated gestures can be applied to social robots with arms. The evaluation result shows the effectiveness of our generative model for speech-driven robot gesture generation.

2.2 Introduction

Gesture as a non-verbal behavior is a supplementary modality of verbal behavior in everyday interpersonal communication. The human can convey meaning and emotion through gesticulation for an expressive interaction. The gesture as a non-verbal body language is very important for a humanoid robot communication during human-robot interaction [33] [13] [141] [142]. Gestures associated and aligned with speech make a robot appear more expressive than a robot using only-speech communication during the interaction [143]. It is very challenging to deal with speech-gesture synchronization problem during the gesture generation with speech [132] [144]. In the past, the natural and human-like robot gestures were mostly handcrafted by researchers again and again, process that is time-consuming and that needs to use prior knowledge in the related domain. Hence, the end-to-end automatic methods of the speech to gesture generation are more and more attracting the attention from researchers. Because the non-verbal behaviors have a random variation to some degree, humans express alternative and variable instead of repeated speech-driven behaviors [145].

The speech-driven gesture for the robot as a non-verbal behavior should possess this kind of characteristic, namely that the robot can express multiple variable gesture patterns given the same speech in different contexts and situations. However, in the past, researchers focused only on one-to-one instead of one-to-many mapping between the speech and the gesture. The work presented in this chapter focuses on the end-to-end speech-driven gesture generation for a humanoid robot. The pipeline of the robot gesture generation is as shown in Figure 2.1. Our 3D gesture synthesizer is based on the Generative Adversarial Network (GAN) [146], capable of using the speech audio as input to generate human-like natural 3D

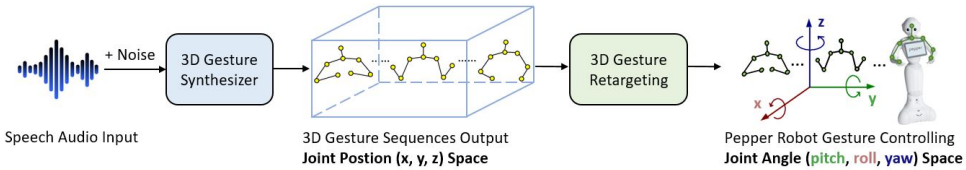


Figure 2.1: The robot gesture generation pipeline. The speech audio is used as an input to the 3D pose synthesizer with a random noise. Then, the natural and human-like 3D gesture sequences (in the joint position(x, y, z) space) are generated. Moreover, one same speech audio with multiple different random noises can align multiple natural gesture expressions in the same manner as humans have similar but different gestures while expressing the same speech in different contexts and situations. The gesture retargeting part maps the obtained gestures in the joint position space of the Pepper robot (i.e., in the joint angle(pitch, roll, yaw) space). Finally, the mapped positions are applied on the real Pepper robot in the human-robot interaction.

gesture sequences, which are in the 3D joint position (x, y, z) space. Then, these 3D positions are mapped to the joint angles (pitch, roll, yaw) of the Pepper robot gesture controlling by our gesture retargeting algorithm. Besides, our model can complete one-to-many mapping work, where one same speech audio with multiple random noises is used as an input to the gesture synthesizer in order to get multiple interrelated gesture sequences.

2.3 State of the Art

2.3.1 Generative model

During the co-speech gesture generation process, the generative model is applied. Hence, first, we will present the state of the art in generative models. Generative models, including the deep belief networks [147], the variational autoencoder models [148], and the generative adversarial models [146] have attracted a lot of attention of researchers from many research areas, including computer vision, robotics, and image processing. These have led to more and more excellent applications of audio-virtual correspondence, where a major focus was on the temporal mapping between the audio and the face action [149] [150] or the human pose [151] [152].

GAN (Generative Adversarial Model) is an effective and popular generation model nowadays. The first GAN model was created by Googfellow et al. [146]. GAN model trains two competitive networks, namely the generator G and the discriminator D in the same time. However, G and D have opposite objectives. G tries to generate the data as the sample in the distribution of the training set, while D tries to differ the real data and the generated data. In a basic GAN model, the G network will be input a random noise sample from a latent distribution such as the normal distribution, and G will produce a sample. The D network is fed with real samples and generated samples from G and output 0 (generated) or 1 (real).

For example, the illustration of the Eiffel tower image generation task with GAN is as shown in Figure 2.3. Input the random noise to the generator to produce the tower image, which is then used as input for the discriminator to be checked as ”real” or “fake“. The real Eiffel tower images are also input to the discriminator during discriminator training. Given random noises as input, the well-trained Generator model can generate new Eiffel tower images never seen in the training dataset.

The Basic GAN above is a kind of unsupervised learning, which has been used in massive image synthesis tasks such as face generation [153] [154]. GAN-based generative model can also be applied to supervised tasks [155]. For example, the GAN model has been used in

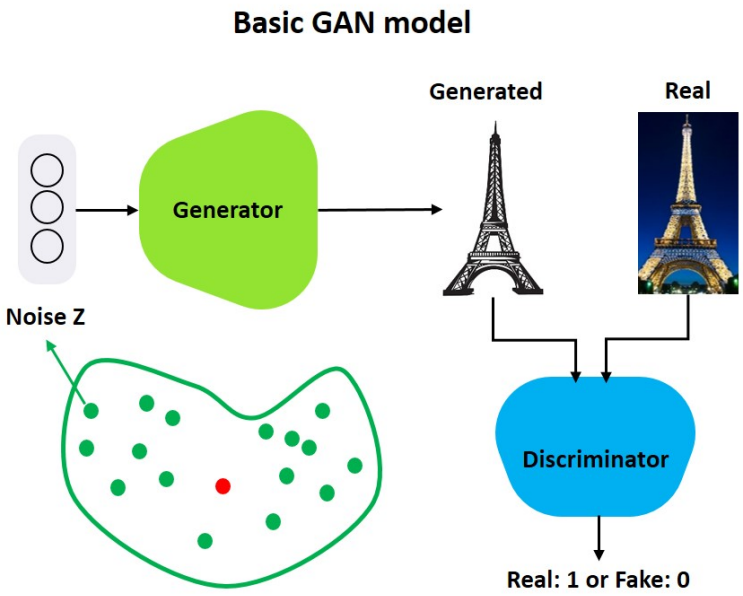


Figure 2.2: The basic GAN model for image generation. GAN contains a generator and a discriminator. The generator (G) tries to produce the samples as the sample in the distribution of the training set. In contrast, the discriminator (D) tries to classify the real samples and generated samples all the time. During the GAN model training process, G and D are trained simultaneously. And G and D contest with each other in this zero-sum game.

many crossmodal generation problems, especially audio-visual mapping. The authors in [150] present an end-to-end co-speech face generation system based on the temporal GAN including a generator and 2 discriminators. The paper [156] utilized a GAN-based model to reconstruct dynamic lip motion given the speech audio as input. The generator of the lip-synchronization networks used the speech's Mel spectrogram features as input to generate the images with lip motion. There are two discriminators. One is a pre-trained discriminator model to check if the lip motions in the generated/real videos match the speech audio input. The other discriminator is a standard CNN-based network to check the visual quality of images. The GAN is finally certified that the synthesized lip motions are natural and aligned based on the human-joined evaluation experiments. In the experiments, the subjects mark the videos containing lip motion and speech based on the sync level, visual quality, preference, and overall experience with audio-visual content together.

The GAN-based crossmodal networks are also used on the music-to-dance synthesis task where the input is the music and the output is the dance actions or images. It is a similar task to the speech-to-face or speech-to-lip tasks. This is because the speech audio and digital music have the same format as a time series, and the generated face/lip video or the dance video are all driven by the regional motion information in images. Hence, the motion generation is the crucial point in these kinds of audio-visual tasks. The authors in [151] developed a music-to-dance model based on GAN to generate the dancing motion sequence (not image) conditioned on the music as input. The model creatively used a synthesis-by-analysis learning framework where a whole dance is decomposed towards a series of basic dance units during the analysis phase. The generative model in the framework also learns how to compose a dance aligned with the given speech audio in the synthesis phase. Then, they generated the realistic and beat-matched dances mapping with the speech.

In addition to the popular GAN model, there are also other generative models or cross-modal mapping models. For example, in [157], the authors presented an RNN-based model

to transfer audio of violin or piano playing to skeletal playing action, including hand movements and upper body movements. MFCC features of speech were extracted as the input, and 2D key points were used for the body actions. Bergmann et al. [158] implemented a Bayesian network to synthesize the iconic gestures. Furthermore, in [159] a co-verbal gesture generation model based on the neural networks and Conditional Random Fields was applied.

2.3.2 Speech-to-gesture generation

Humans expect the social robot to respond with similar manners as humans do during human-robot interaction. Interpersonal communication refers to many types of human behaviors, including verbal behaviors and non-verbal behaviors. The robot also should have these kinds of behaviors with the help of some behavior generation models. Robot verbal behaviors relate to the speech with text and vocalization, which are necessary for a social robot to convey meanings directly. Robot non-verbal behaviors also play an essential role in natural human-robot interaction. The gesture is a crucial component to share the meanings and feelings during non-verbal communication, which may refer to hand action, arm action, head action, and body action.

The gesture can be divided into four types: beat gestures, iconic gestures, metaphorical gestures, and deictic gestures [160]. An iconic gesture represents object attributes, spatial relationships, and actions. For example, you draw a curve to describe a mountain. That gesture is an iconic one to indicate a concrete object during communication. Instead of referring to a concrete thing, the metaphorical gesture is related to the abstract content. For example, you shape your hand with the heart shape and place it on your chest to convey love to others. The deictic gesture (or pointing gesture) often refers to the direction or position of the object described by simultaneous verbal speech. For example, when you start to talk about your audiences, you use the deictic gesture to point towards the people. A beat gesture is a gesture simultaneous with the speech rhythm. That kind of rhythmic gesture cannot convey the semantic information but can emphasize a specific word or sentence for expressive communication. Among four types of gestures, the beat gesture and iconic gesture are often studied when the researchers work on the gesture synthesis [161] [162] [163]. The beat-based gesture generation refers to the simultaneous speech utterance, namely speech-audio-driven gesture generation. The iconic gesture generation is related to verbal speech, namely, speech-text-driven gesture generation. Hence, we will first discuss the related works with three gesture generation methods based on the type of input speech, namely the speech with text, the speech input with audio and the speech input with both.

Speech-driven Gesture Generation with Text

The verbal speech can convey more direct meanings during communication. The gesturing with speech text can make a more expressive communication for an embodied agent. The text-driven gesture generation model learn a mapping from semantic features of speech to gestures. The early work on text-driven gesture generation used the rule-based model to generate the co-speech gestures. For example, the authors in [164] built-up a ruled-based method to correlate the speech text with the hand-crafted gesture. Aly et al. [165] built a generation model with the tool BEAT (Behavior Expression Animation Toolkit) [166] to map a human's verbal behavior to a corresponding combined robot's verbal-nonverbal behavior, namely gesture, based on the personality dimensions of the interacting human. However, the BEAT toolkit is a rule-based method and generates repetitive gestures without a random variation. With the development of deep learning, more and more researchers focus on AI-based crossmodal mapping method for text-driven gesture generation. Yoon et al. [167] built

a large-scale co-speech gesture database including speech text and spontaneous 2D human gestures extracted from public videos of Youtube TED talks. The authors proposed an autoencoder model based on the *seq2seq* model [24] and the generative model used the speech text as the input to generate 2D human gesture sequence. The authors also applied the generated pose and the speech audio synthesized from text to the NAO robot for the participant evaluation experiments. However, as mentioned in their work, the method experienced an unnatural mapping problem where the generated gesture and the speech audio could not be tightly mapped well during the experiments. The synchronization problem is a challenge for all the text-based gesture synthesis model as they do not take the acoustic features of speech into consideration during gesture generation.

For the same text sentence, different people can speak with variable emotional states and speeds, which should correspond to different poses. However, this speech text-based approach can align only one pose pattern with only one speech sentence, which leads to an unnatural and not human-like human-robot interaction. Speech audio-based method can deal with this well in order to generate different pose sequences based on the variable speech expression of speakers.

Speech-driven Gesture Generation with Audio

Essentially, the audio-based gesture generation is driven by the prosody of speech input. Audio data is less structured compared with the speech text and thus harder to model in crossmodal mapping task [151]. Hasegawa et al. [168] used the Bi-Directional LSTM (Long Short-Term Memory) Network to generate the co-speech gestures. The model processed the speech audio into MFCC (Mel-Frequency Cepstral Coefficients) as the audio representation and also included LSTM regression part for the temporal mapping and a temporal filtering to remove noises of the 3D pose data. Using the same database, Kucherenko et al. [169] proposed an autoencoder-based approach to deal with the alignment between the speech audio and the human pose. One autoencoder network, namely *MotionED*, was trained for lower dimensional motion representation and one encoder was trained for the mapping the speech to motion representation obtained in *MotionED*. In [170], a speech and gesture generation model using the text as input was described. The authors used a Tacotron-based text-to-speech (TTS) model [27] to generate the spontaneous speech utterances given texts as input. The generated speech was fed to a probabilistic model, namely MoGlow [171], to generate the aligned gesture. Essentially, it is a speech-audio-driven gesture generation. The different part is that the paper used the synthesized speech from text instead of natural human speech as input during speaking gesture generation.

Instead of generating the same gesture from one input speech, the authors in [171] built up a probabilistic generative sequence model based on a normalising flow model to produce the style-controllable gesture from speech audio. The controllable style contains the gesture level, speed, symmetry, and spatial extent, leading to an adaptive and natural human-robot interaction. The paper [172] applied a GAN-based model to synthesize the speaking gestures with different styles referring to different persons. The generator of GAN was fed with MFCC features of speech audio to produce the related aligned gesture sequence. The discriminator of GAN tried to ensure that the predicted motion is temporally coherent and in the speaker's style. This work inspires many crossmodal mapping tasks from audio to visual features.

Speech-driven Gesture Generation with text and audio

The authors in [32] came up with a text-driven gesture generation system for a humanoid robot with arms and head. Authors used a WordNet tool to represent the similar words into

the feature space and the similar word embedding was mapped to a similar gesture mode with a probability-based model. The paper also explored a prosody-driven beat gesture generation system, which tried to map the prosodic peaks of the fundamental frequency speech signal to the beat gesture. In other words, if the prosodic peak is detected, the synthesis system will produce a beat gesture in that position with robot hands down and up about 15 centimeters from the current gesture. The prosodic gesture generation also used in the paper [173]. Finally, the two kinds of generated gesture signals were overlaid together for the humanoid robot during HRI. However, in this system, the text-driven gesture and prosody-driven gesture are synthesized independently, which is challenging to get natural and synchronous multimodal HRI.

Differently, Kucherenko et al. [174] built-up an RNN-based model, which can generate beat and iconic gesture together. The model took both acoustic (audio) and semantic (text) representations to synthesize synchronous gestures with joint angle rotations, which can be used for an avatar or humanoid robot. Yoon et al. [31] proposed a gesture synthesis model that took multimodal information with text, audio, and speech identity to produce the human-like gesture. The proposed model can generate human-like gestures mapping with the speech content and speech rhythm by integrating the trimodal features.

2.4 Our Methodology

2.4.1 Problem definition

Speech-driven Gesture Generation: This problem is an one-to-many task by nature to generate the diverse spontaneous gesture sequences $g^m = [g_t^m]_{t=1:T_0}$ with one speech audio $s = [s_t]_{t=1:T_1}$ and multiple noises n^m as inputs. Namely, the research tried to learn a mapping function $F_{generation}$, which will maximize the conditional probability $p(g^m|s)$.

$$\mathbf{g}^m = F_{generation}\left(\mathbf{s}, \mathbf{n}^m\right) \tag{2.1}$$

3D Gesture Retargeting: The objective is to map the generated gestures g^m with the position $p(x,y,z)$ in the joint position space towards the robot gestures r^m with the angle $a(pitch, roll, yaw)$ in the joint angle space, which can be used on the Pepper robot co-speech gesture controlling directly. The mapping consists in finding a retargeting function $F_{retargeting}$:

$$\mathbf{r}^m = F_{retargeting}\left(\mathbf{g}^m\right) \tag{2.2}$$

2.4.2 Gesture generation model

A speech refers to many natural gestures that humans can conduct during speech. The mapping between speech and gesture sequence is weaker than image-to-image translation, such as semantic segmentation as a rigid one-to-one mapping. However, the traditional crossmodal mapping models (such as RNN) utilize the frame-wise or pixel-wise loss during generative model training, which often suffer the known issue of regression to the mean, which produces overly smooth motion [172]. The adversarial loss of GAN considers whether the whole gesture sequence match with the speech input instead of a frame-wise rigorous mapping. This is the reason why GAN is used in our work.

In order to facilitate the co-speech gestures task, we built a new temporal generative adversarial network (GAN), which contains the generator and the discriminator. In this thesis, the proposed GAN model as shown in Figure 2.3 can produce diverse speech-matching gestures, which can be applied towards the humanoid robot. Our GAN model is named with *S2GGAN*, namely **S**peech to (**2**) **G**esture **GAN**. *S2GGAN* model contains a generator and a discriminator. The generator includes one encoder and one decoder and takes the speech

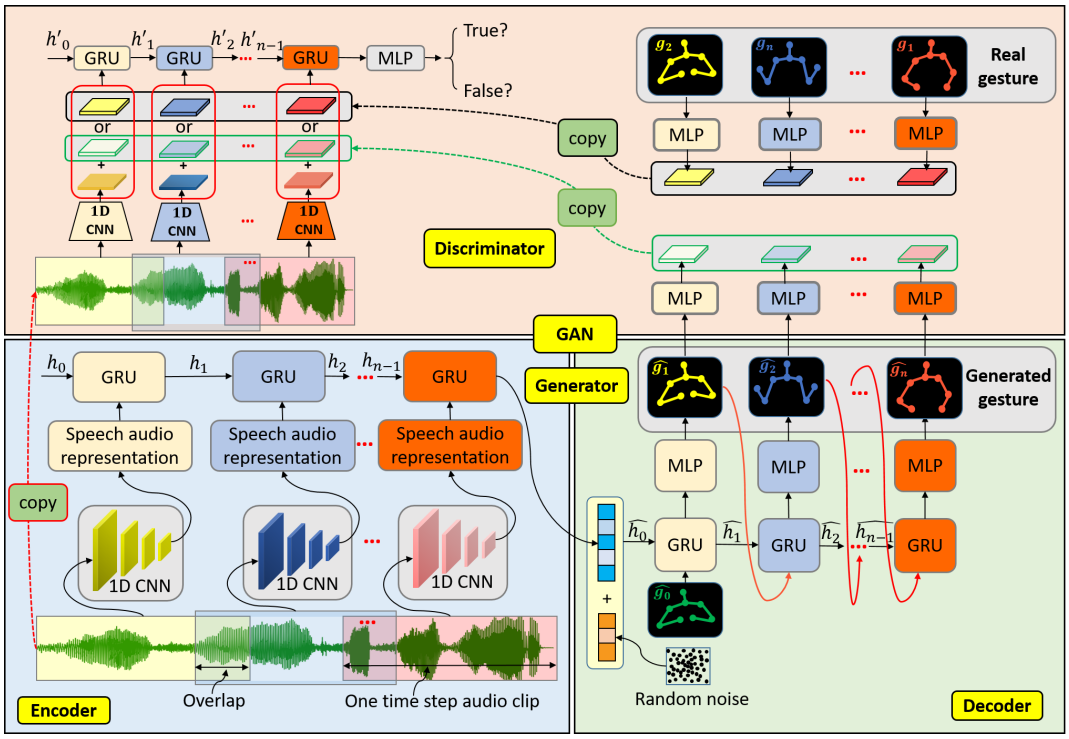


Figure 2.3: An overview of our GAN model architecture for speech-driven gesture generation. The whole GAN model consists of a generator and a discriminator. The generator contains a temporal encoder and a temporal decoder. The encoder takes the speech audio as input to get the last hidden state as output for the later decoder input. The next decoder is applied to decode the input with the encoder output and a random noise towards the mapping gesture. The discriminator uses the generated gesture (or the ground-truth gesture) and the spontaneous speech audio as input to predict whether the speech and the gesture match with each other.

audio as input and outputs the mapping gesture. The discriminator is used to predict whether the speech and the gesture match with each other.

Generator

The generator consists of an encoder and a decoder and uses the speech audio as input and outputs the spontaneous gesture sequence extracted from YouTube TED videos. As shown in Figure 2.3, the audio signal is divided into overlapping audio clips so that each audio clip is one-to-one correspondence with the gesture frames.

Then, the audio clips should be represented as features with useful information guiding gesture generation. Primarily, researchers use Mel-frequency cepstral coefficients (MFCCs), and 1D CNN (Convolutional Neural Network) for speech audio representation [12] [175] [150] [176]. MFCC is a frequency-based feature of audio, which focuses on acoustic prosody, which is training-free. Except for the deeper acoustic features compared to handcrafted MFCC, 1D CNN also can extract the semantic representation to some degree. In this chapter, 1D CNN is used to extract the speech audio representation. 1D CNN can extract the 256 dimensional representation for each audio clip. 1D CNN structure [150] is applied in our encoder network and it is composed of 1D convolution, a batch normalization part, and a ReLU part. The 1D CNN structure is as shown in Figure 2.4. Inspired from [177], the 1D CNN should start with a a large kernel, which can make the low-level feature meaningful. In our work, the first kernel size is 250 and all the followed kernel sizes are 4. We chose leaky

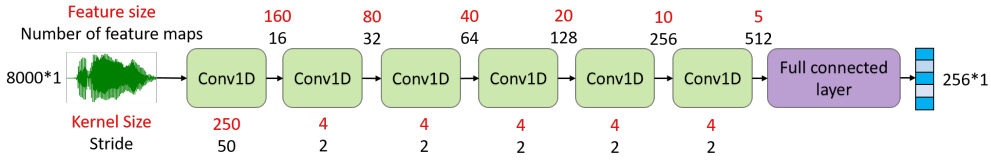


Figure 2.4: 1D CNN network framework. 1D CNN model input is 8000 frame audio clip and output is 256 dimensional representation. This 1D CNN start with a big kernel,namely 250 and all the followed convolution operations have a small kernel size with 4. Among convolution operations, there are leaky ReLU operation and batch normalization operation. Finally, a full connected layer is used to get the 256 dimensional speech audio representation.[150]

ReLU activation [178] instead of a common ReLU because the latter experiences a dying ReLU problem and the former can overcome this problem during the training. Lastly, these representation results are fed into Gated Recurrent Unit (GRU) model [15] to get the 256 dimensional final hidden state as the encoding output of the encoder network.

The decoder network comprises one layer GRU and the Multilayer Perceptron (MLP). The input of the decoder is the output of the encoder adding the 10 dimensional random noise. As described in [179], the random noise can introduce some natural variability during the audio-visual generation task. We used multiple random noises to generate various kinds of mapping gestures from one speech audio. This can produce various gestures for the robot and can lead to a natural long-life human-robot interaction. The output of the decoder is the spontaneous gesture sequence mapping with the speech audio in the encoder network. Each gesture frame contains 3D positions of 8 joints including the head, the spine shoulder joint, the left shoulder joint, the right shoulder joint, the left elbow joint, the right elbow joint, the left wrist joint, and the right wrist joint.

Discriminator

The discriminator works to distinguish whether the gesture sequence is realistic and whether the gesture sequence matches with the speech audio or not. Firstly, the 1D CNN, which is the same with the one in the generator, receives the speech audio divided into the overlapping clips as one of two inputs to get the 256 dimensional representation vectors. Meanwhile, each gesture frame of the real or generated gesture sequence is used as an input to the MLP layer to get 256 dimensional vectors. These vectors concatenate 256 dimensional speech representation vectors each by each respectively to get the 512 dimensional vectors as the input of one layer GRU. The GRU final hidden state is used as input for the MLP layer to distinguish whether the gesture matches with the speech audio.

Objective function

The objective function of our *S2GGAN* contains two parts, namely conditional GAN loss part and L_1 loss part. The conditional GAN loss part can be expressed as shown in Equation 2.3. In this equation, G tries to minimize the loss while D tries to maximize it.

$$\begin{aligned}\mathscr{L}_{cGAN}(G,D) =& \mathbb{E}_{s,g}[\log D(s,g)]+ \\ & \mathbb{E}_{s,z}[\log(1-D(s,G(s,z))]\end{aligned} \tag{2.3}$$

where s denotes the speech audio, g means the gesture sequence, and z is the random noise.

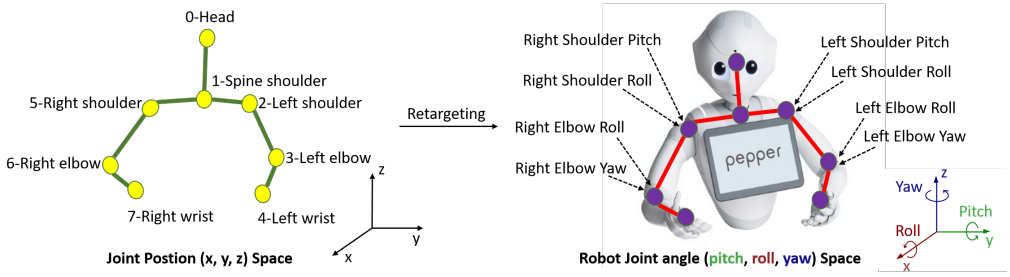


Figure 2.5: An overview of the gesture retargeting process. In the joint position (x,y,z) space, there are 8 joints' 3D positions. The joints are the head (joint 0), the spine shoulder joint (joint 1), the left shoulder joint (joint 2), the right shoulder joint (joint 5), the left elbow joint (joint 3), the right elbow joint (joint 6), the left wrist joint (joint 4), and the right wrist joint (joint 7). Based on the robot kinematics, the 8 angles of the 4 joints are obtained from the 3D positions. The 8 joint angles are the left shoulder pitch, the right shoulder pitch, the left shoulder roll, the right shoulder roll, the left elbow roll, the right elbow roll, the left elbow yaw, and the right elbow yaw. Joint roll rotations take place around the X axis, joint pitch rotations around the Y axis, and joint yaw rotations around the Z axis.

Previous works have found that it is beneficial to add L_1 loss towards the conditional GAN loss in image translation task [180]. Moreover, using the GAN loss with limited database is challenging to train GAN model. In our work, the L_1 reconstruction loss is applied to enhance the realistic gesture generation at the frame level (see Equation 2.4).

$$\mathscr{L}_{L1}(G) = \mathbb{E}_{s,z,g}\left[\|g - G(s,z)\|_1\right] \tag{2.4}$$

The final objective of our *S2GGAN* is as shown in Equation 2.5. The hyperparameter λ is empirically set to 100 during *S2GGAN* training, which is used to decide how much contribution L_1 or L_{cGAN} make for all the loss.

$$G^*_{S2GGAN} = \arg\min_G \max_D \mathscr{L}_{cGAN}(G,D) + \lambda \mathscr{L}_{L1}(G) \tag{2.5}$$

2.4.3 Gesture retargeting

Mapping from 3D joint positions $p(x,y,z)$ to 3D robot joint angles $a(roll, pitch, yaw)$ is a problem of robot kinematics. There are a lot of works showing how to transfer human pose with 3D positions to robot pose with joint motor angles [181] [182]. In our work, we use Pepper robot to present the generated co-speech gesture. The Pepper is a social humanoid robot from SoftBank Robotics. In our work, 8 joint positions are generated, including the head, the spine shoulder joint, the left shoulder joint, the right shoulder joint, the left elbow joint, the right elbow joint, the left wrist joint, and the right wrist joint. Furthermore, 8 joint rotation angles of 4 joints should be obtained from the 3D positions in each gesture frame. These 8 joint angles of the robot upper body gesture can be obtained from the 3D positions of the 8 joints as shown in Figure 2.5.

The definitions and the angle ranges of the 8 joint rotation angles are shown in Figure 2.6 according to the Pepper robot official technique documents [183]. In order to make sure that the rotation axes in the left shoulder joint are in the rotation angle space, each unit vector along each positive direction of x axis, y axis, and z axis, respectively should be determined. These three unit vectors are defined as shown in Equations 2.6, 2.7, and 2.8, respectively.

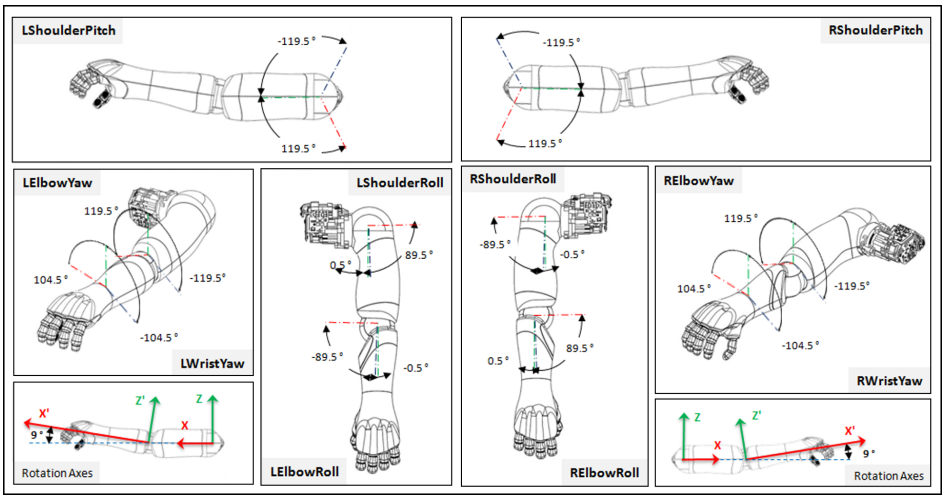


Figure 2.6: The definition of the joint rotation angles and their rotation angle ranges in degree. [183]

$$\vec{x} = \frac{\vec{V}_{0,1} \times \vec{V}_{1,2}}{|\vec{V}_{0,1} \times \vec{V}_{1,2}|} \tag{2.6}$$

$$\vec{y} = \frac{\vec{V}_{1,2}}{|\vec{V}_{1,2}|} \tag{2.7}$$

$$\vec{z} = \frac{\vec{x} \times \vec{y}}{|\vec{x} \times \vec{y}|} \tag{2.8}$$

where $\vec{V}_{m,n}$ is a vector from the position of the joint m to the position of the joint n $(m,n = 0,1,2,3,4,5,6,7)$. The relation between the joint number and the joint name is as shown in Figure 2.5. In the mechanical structure of the physical Pepper robot, the spine shoulder joint, the left shoulder joint, and the right shoulder joint keep the fixed spatial relationship. In our research, we fix the position spine shoulder joint towards the position of midpoint of the left shoulder joint and the right shoulder joint.

For the left and right shoulder joints, there are 4 joint rotation angles that should be extracted, including the left shoulder roll, the left shoulder roll, the left shoulder pitch, and the right shoulder pitch.

In order to simplify the calculation process, we used the unit vectors of the left shoulder joint rotation axes, namely $\vec{x}$, $\vec{y}$ and $\vec{z}$, for these 4 angles calculation of both the left shoulder and the right shoulder joint. *LSR* (Left Shoulder Roll), *RSR* (Right Shoulder Roll), *LSP* (Left Shoulder Pitch) and *RSP* (Right Shoulder Pitch) are calculated as shown in Equations 2.9, 2.10, 2.11, and 2.12, respectively.

$$LSR = \frac{\pi}{2} - \csc^{-1}\left(\frac{\vec{y} \cdot \vec{V}_{5,6}}{|\vec{y}| \cdot |\vec{V}_{5,6}|}\right) \tag{2.9}$$

$$RSR = \csc^{-1}\left(\frac{\vec{y} \cdot \vec{V}_{2,3}}{|\vec{y}| \cdot |\vec{V}_{2,3}|}\right) - \frac{\pi}{2} \tag{2.10}$$

$$LSP = -\tan^{-1}\left(\frac{\vec{V}_{2,3} \cdot \vec{z}}{\vec{V}_{2,3} \cdot \vec{x}}\right) \tag{2.11}$$

$$RSP = -\tan^{-1}\left(\frac{\vec{V}_{5,6}\cdot\vec{z}}{\vec{V}_{5,6}\cdot\vec{x}}\right) \tag{2.12}$$

For the left and right elbow joint rotation angles, their related coordinate axes rotate because of the shoulder joints' rotation. Relative to the calculation of four elbow rotation angles, only the unit vectors of x axis should be transferred, including $\vec{x}_{left}$, and $\vec{x}_{right}$. The related calculation of the 4 elbow joint rotation angles are shown in the Equations 2.13 - 2.19. Where $\vec{v}_{\text{plane234}}$ is the normal vector of the plane defined by joint 2 (left shoulder), joint 3 (left elbow) and joint 4 (left wrist). $s(\alpha)$ and $c(\alpha)$ means the sine and cosine of α respectively. LER, RER, LEY, and REY are the left elbow roll angle, the right elbow roll angle, the left elbow yaw angle and the right elbow yaw angle, respectively.

$$LER = -\cos^{-1}\left(\frac{\vec{V}_{2,3}\cdot\vec{V}_{3,4}}{\left|\vec{V}_{2,3}\right|\cdot\left|\vec{V}_{3,4}\right|}\right) \tag{2.13}$$

$$RER = \cos^{-1}\left(\frac{\vec{V}_{5,6}\cdot\vec{V}_{6,7}}{\left|\vec{V}_{5,6}\right|\cdot\left|\vec{V}_{6,7}\right|}\right) \tag{2.14}$$

$$\begin{aligned}\vec{x}_{left} = &\begin{bmatrix} c(LSR) & s(LSR) & 0 \\ -s(LSR) & c(LSR) & 0 \\ 0 & 0 & 1 \end{bmatrix}\cdot \\ &\begin{bmatrix} c(LSP) & 0 & s(LSP) \\ 0 & 1 & 0 \\ -s(LSP) & 0 & c(LSR) \end{bmatrix}\cdot\vec{x}\end{aligned} \tag{2.15}$$

$$\vec{v}_{\text{plane234}} = \frac{\vec{V}_{2,3}\times\vec{V}_{3,4}}{\left|\vec{V}_{2,3}\times\vec{V}_{3,4}\right|} \tag{2.16}$$

$$LEY = \cos^{-1}\left(\frac{\vec{x}_{left}\cdot\vec{v}_{plane234}}{\left|\vec{x}_{left}\right|\cdot\left|\vec{v}_{plane234}\right|}\right) - \frac{\pi}{2} \tag{2.17}$$

$$\begin{aligned}\vec{x}_{right} = &\begin{bmatrix} c(RSR) & s(RSR) & 0 \\ -s(RSR) & c(RSR) & 0 \\ 0 & 0 & 1 \end{bmatrix}\cdot \\ &\begin{bmatrix} c(RSP) & 0 & s(RSP) \\ 0 & 1 & 0 \\ -s(RSP) & 0 & c(RSR) \end{bmatrix}\cdot\vec{x}\end{aligned} \tag{2.18}$$

$$\vec{v}_{\text{plane567}} = \frac{\vec{V}_{5,6}\times\vec{V}_{6,7}}{\left|\vec{V}_{5,6}\times\vec{V}_{6,7}\right|} \tag{2.19}$$

$$REY = \cos^{-1}\left(\frac{\vec{x}_{right}\cdot\vec{v}_{plane567}}{\left|\vec{x}_{right}\right|\cdot\left|\vec{v}_{plane567}\right|}\right) - \frac{\pi}{2} \tag{2.20}$$

In addition, after gesture retargeting from generated human gesture to robot gesture, the median filter was used to remove the sudden jittering. The sudden jittering often leads to an unacceptably high rotation speed for the robot joint motor. And then, the Pepper robot will stop to conduct the gesture with a motor speed more than the maximum to protect the joint motors.

Figure 2.7: TED speaking scene. The images extracted from the YouTube TED videos

2.5 Experiments and results

2.5.1 Database building

It is tough to build-up a vast audio-visual database with speech and spontaneous gestures. Recently, instead of the hand-crafted database with limited size from the lab, some researchers tried to make full use of the well-made public videos that recorded multimodal audio-visual information naturally, for example, the pubic Youtube videos. There are many advancements to build the database on a public video collection. Firstly, enough well-recorded videos with clear audio and high-definition images are available online, which makes database building time-saving and money-saving. The video database also grows with time. Oh et al. [184] millions of natural Internet/YouTube videos of people speaking for speech-driven face generation task. Yoon et al. [167] proposed a solution to build-up a big speech-driven gesture database; they used 52 hours of public YouTube TED videos and related English transcripts. Using *OpenPose* [185], which is a popular Human Pose Estimation (open-source) library, 2D human poses were extracted from videos. Then, they built the 2D speech driven gesture with the speech text and the synchronous 2D pose. Inspired by these works, we built up our own speech-driven database with the speech audios and the spontaneous 3D gestures extracted from YouTube TED videos. The TED speaking scene can be seen in the Figure 2.7 Instead of speech text, we used speech audio in our research as the same speech text can be expressed as various speech audios. Instead of 2D gesture, we achieved 3D gesture from 2D gesture. The whole process of our speech-driven gesture database building is as follows.

(1) Download public TED videos from YouTube through *YouTube Data API*.

(2) Extract the speech audios from videos downloaded through the $FFmpeg$ library [186].

(3) Extract 2D gesture data of eight joints used in our study from YouTube videos through *OpenPose* library.

(4) Transform the 2D gesture to the 3D gesture of eight joints by $3D-pose-baseline$ model trained by ourselves.

(5) Based on the rules of clip selection, we divided the data into clips and built the database for speech-driven gesture generation task.

Audio pre-processing

The speech audios are extracted from TED videos through $FFmpeg$ interface, and the obtained audio is stereophonic with the left channel and right channel. However, only mono track is needed for the generative model. During the transformation from stereo audio to mono audio, we applied the mean values of the left channel and right channel as the mono channel audio for the generative model, as shown in Figure 2.8, which method was also used in many other audio-related process tasks, e.g., in [169].

The speech audio data and gesture data have different frame rates. The audio frame rate is 44100 Hz, and the gesture frame is 24 Hz (same with the selected videos). In the

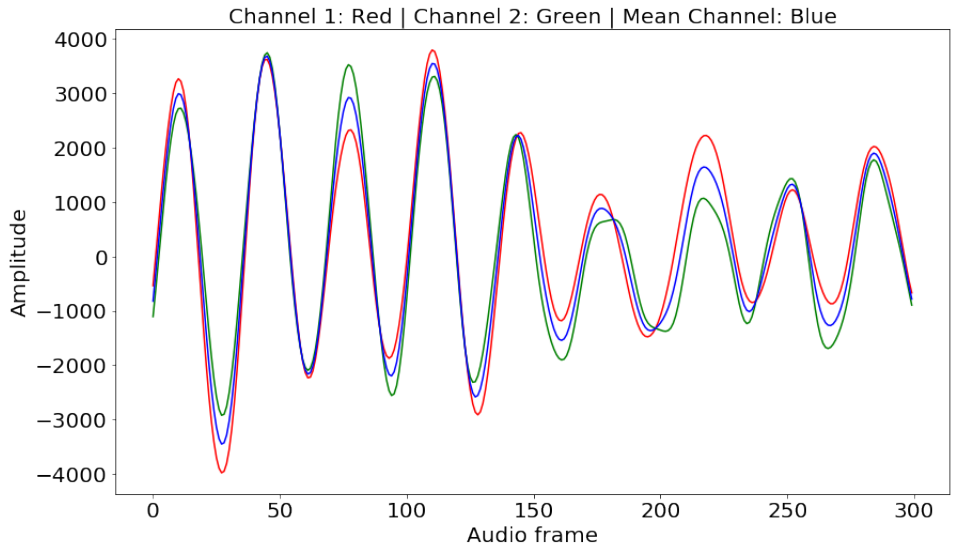


Figure 2.8: The mono audio extraction from the stereo audio. The illustration only visualizes one audio clip with 300 frames as an example.

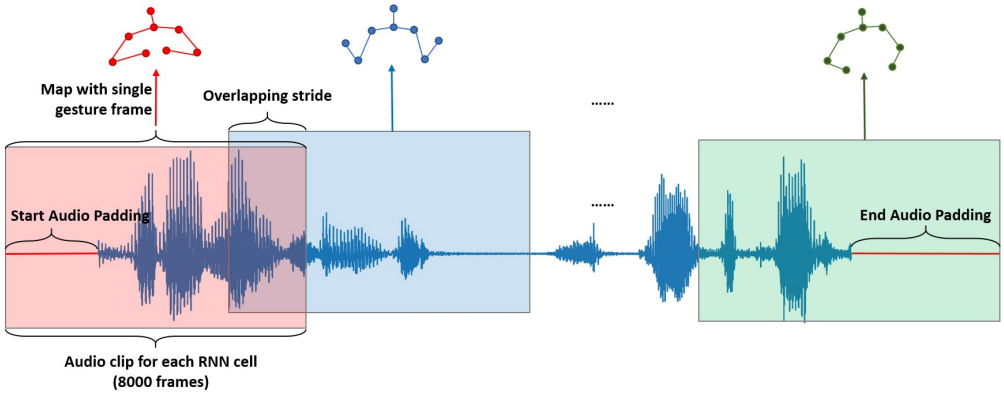


Figure 2.9: Audio clip cutting and overlapping stride.

strict sense, one gesture frame should correspond to 1837.5 audio frames. However, the generator's encoder with an RNN network (GRU) is fed with a more extended audio overlapping clip and outputs a single frame at each time step. The overlapping audio clips are cut from the whole audio sample as shown in Figure 2.9. The overlapping clips are used instead of non-overlapping ones because gestures can sometimes lag behind or rush ahead of your speaking contents. For example, when you think considerably during interpersonal communication, your gesture will run faster than your speech. However, when you have well-prepared speaking content, for example, in a public speaking scene, your gesture will lag behind your speech. Hence, each audio clip for each time step of the following GRU network contains two continuous parts, including the past clip and future clip in order. Each audio clip has 8000 audio frames in total. Namely, there are 4000 past audio frames and 4000 future audio frames selected around the time point of the relative gesture frame. Therefore, in order to deal with alignment between the audio clip and the gesture frame, the starting part and the ending part of the whole audio signal are padded with zeros, respectively.

The audio clip stride is around the quotient of the audio frame rate (44100 Hz) divided by the gesture frame rate (24 Hz) to keep the temporal alignment between speech audio clip and gesture frame. Because this quotient (1837.5) is a non-integer, the stride used at each

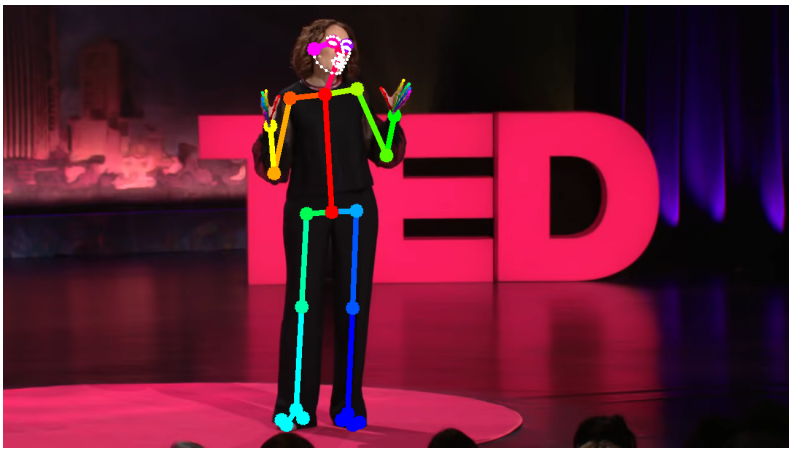


Figure 2.10: *OpenPose* for whole-body estimation.

time step is calculated as the following equations.

$$f_{\text{ratio}} = \left\lfloor \frac{f_{\text{audio}}}{f_{\text{gesture}}} \right\rfloor \tag{2.21}$$

$$\text{Stride}\ (ts) = \begin{cases} f_{\text{ratio}} + 1, & \text{ts} = \text{even number;} \\ f_{\text{ratio}} - 1, & \text{ts} = \text{odd number.} \end{cases} \tag{2.22}$$

Where, f_{ratio} is the floor function of the quotient of the audio frame rate divided by the gesture frame rate. ts id the time step of gesture sequence. Stride(ts) is the stride at the time step ts.

2D Gesture extraction

During database building, 2D gesture frames are extracted from the image frames of TED videos through the *OpenPose* interface. *OpenPose* is a popular open pose estimation library that can detect whole-body 2D poses with body, foot, face, and hands, as shown in Figure 2.10. In our work, we only focus on the upper body gesture, and the used eight joints are the head, the spine shoulder joint, the left shoulder joint, the right shoulder joint, the left elbow joint, the right elbow joint, the left wrist joint, and the right wrist joint.

There are 1760 downloaded from the public YouTube TED collections. Considering the synchronization of the speech audio and the gesture in our *S2GGAN* model, we only selected the videos with a frame rate of 24 Hz (most of the videos are 24 Hz and few of them are 25 Hz). Then, the extracted pose sequences are cut into pose clips. Not all the detected pose clips can be used for the generative model training. For example, the speaker stands with the back; the audiences are detected in one frame; we cannot detect the whole upper body; there are more than one speakers; the video scene is not with the natural speaking gestures, and so on. The not suitable scenes for the database are as shown in Figure 2.11. During the database building process, we made our rules for the continuous clip selection based on the rules discussed in the paper [167]. Our gesture clip selection rule as follows:

- The eight upper body joints are detected in all frames of the gesture clip.
- The speaker in the selected clip should face the camera.
- The gesture clips are more than 5 seconds.
- The TED speaking scene only contains one speaker.

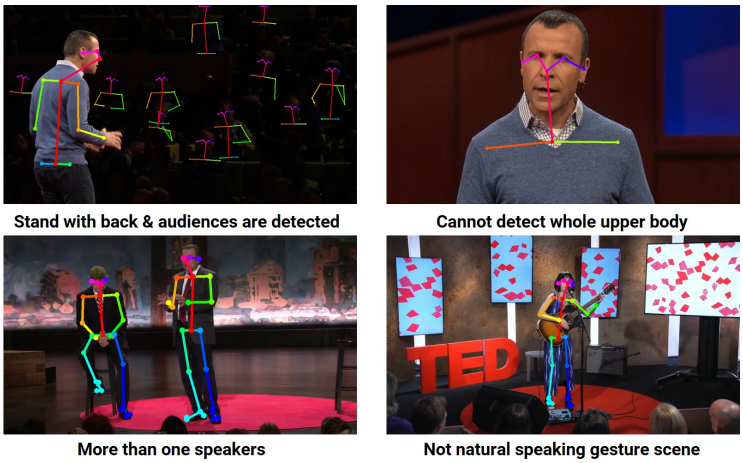


Figure 2.11: *OpenPose* examples with wrong results.

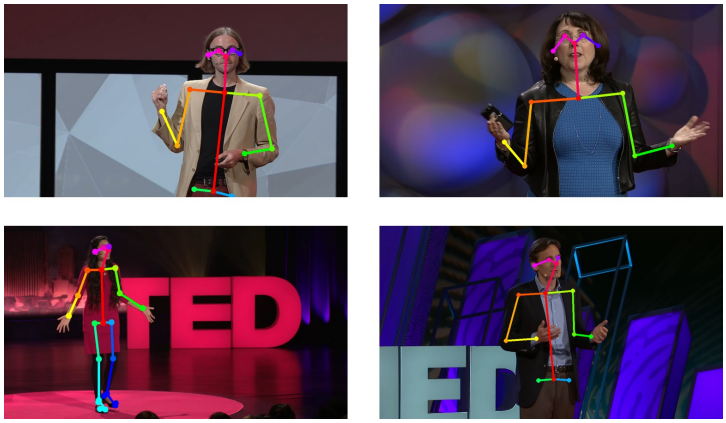


Figure 2.12: *OpenPose* examples with suitable results.

- No audiences are detected in the selected gesture clips.
- The video scenes is with the natural speaking gestures.
- There should be no still frames where the speaker stays still without the gesture movements.

The gesture clip selection process with these rules is conducted with the automatic programming method and handcrafted method. Even so, the errors are unavoidable, which will be possibly covered by most of the suitable gesture clips during training. The gesture samples selected for database building are as shown in Figure 2.12. Then those selected gesture clips are divided into shorter clips with the same duration. Moreover, only eight joints' data are chosen for the database.

3D Gesture extraction

Because we need a 3D pose for real robot conducting gesture, the 2D gestures are transferred to 3D gesture through the $3D-pose-baseline$ library [187]. $3D-pose-baseline$ is a lightweight and effective library doing 2D-to-3D pose estimation, and its code is open for public usage. The pipeline is as shown in Figure 2.13. The pose lifting model based on the neural networks is fed with 2D positions of whole-body joints and outputs the 3D positions. The model processes the pose frame by frame. It runs effectively and fast during training and testing.

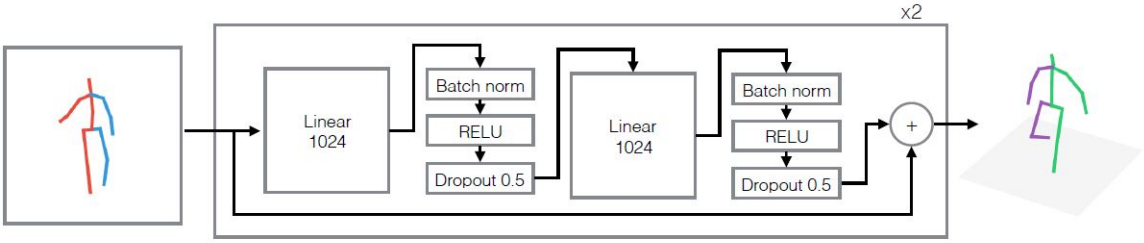


Figure 2.13: $3D-pose-baseline$ pipeline [187]

Figure 2.14: 3D position plot of left wrist joint.

However, the original model was trained for the whole body pose data. Our work focuses only on the upper body gesture because the whole body cannot be seen in most YouTube TED videos. Hence, to complete our 2D-to-3D pose estimation task, we trained the $3D-pose-baseline$ with eight joints data from the $Human3.6M$ dataset [188], which is an open 3D human pose database with 3.6 million human poses and corresponding images. Then, the trained model is used on later 3D pose extraction for our database. The 3D position sequences of wrist joints randomly sampled from the database are visualized as shown in Figure 2.14 and Figure 2.15. In addition, the median filter (window size = 5) is used on each gesture time series for denoising. Finally, the database for speaking gesture generation contains 5760 samples in total with speech audio clips and 3D gesture clips.

2.5.2 Model training

Our audio-gesture database contains 5760 samples, which were divided into 180 batches with the batch size of 32. The training size : validation size : testing size is 135 batches : 15 batches : 30 batches. The time steps of the gesture sequence were fixed to 126 and the time duration is of 5.25 seconds. Standardization operation is completed on the 3D gesture data before inputting *S2GGAN* model and batch normalization operation was used in 1D CNN part of *S2GGAN*, which both can effectively reduce the over fitting during model training based on our tuning experiments. We used Adam [189] as the optimization algorithm where the learning rate is 0.0002, the parameter β_1, β_2 and ε are 0.5, 0.999, and 10^{-7}, respectively for both the generator and the discriminator in *S2GGAN* model. Our *S2GGAN* model is

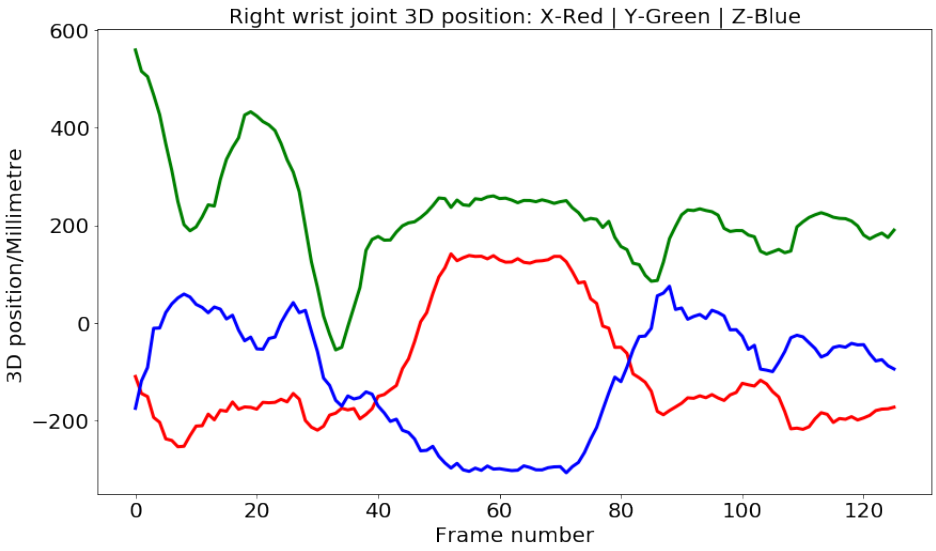


Figure 2.15: 3D position plot of right wrist joint.

completed in TensorFlow 2.0 and the training process with 4086 epochs was done on an NVIDIA GeForce RTX 2080 Ti GPU for about 1 week. The discriminator losses (real and generated), generator loss, and $L1$ loss during the training process are shown in Figure 2.16. The generator and discriminator converge from about 1000 epochs while the $L1$ loss still keeps declining.

2.5.3 Results

After *S2GGAN* training, we can test the model to generate various speech-driven gesture sequences with one speech audio and multiple random noises as inputs. Here, we used two noises to generate two gesture sequences for each speech audio. After gesture generation, the denormalization operation was performed on the synthesized 3D gestures because the gesture data was normalized previously during the training. Then, the robot gesture can be run on the Pepper robot after retargeting from the joint positions to the joint rotation angles. Pepper is a semi-humanoid robot manufactured by SoftBank Robotics. It has flexible arms, which can conduct human-like gestures. The synthesized robot gesture can work on the real Pepper robot or in a simulation environment-*Choregraphe*. The real Pepper robot and the visual Pepper robot are shown in Figure 2.17. We applied the joint rotation angels to run the virtual Pepper robot in *Choregraphe* to check the generated speaking gestures. [1].

The generated gestures based on audio speech were used to make the Pepper robot conduct multimodal behaviors with gesture and speech audio, which makes a social robot more expressive than a robot using only speech during human-robot interaction. One example is shown in Figure 2.18 where the sampling rate is 5 fps (frames per second) and gesture duration is 5 seconds. The related speech text is "I removed my timing chip, and I handed it over to a race official." The generated robot gesture looks expressive. Related video can be seen in Gesture Video Link.

Qualitative evaluation

As any one-to-many generative model, our *S2GGAN* is able to generate multiple gesture sequences mapping with one same speech audio and multiple random noises. After pose

[1]Due to Covid19 and the multiple lockdown and restrictions, we could not use the real robot to run the experiments

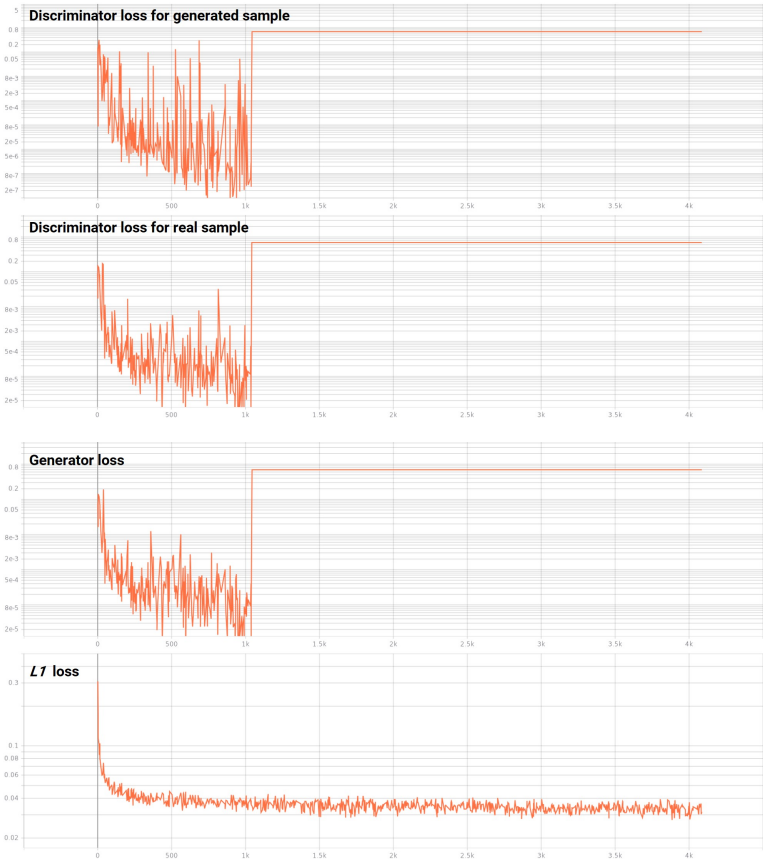


Figure 2.16: GAN losses plot during training.

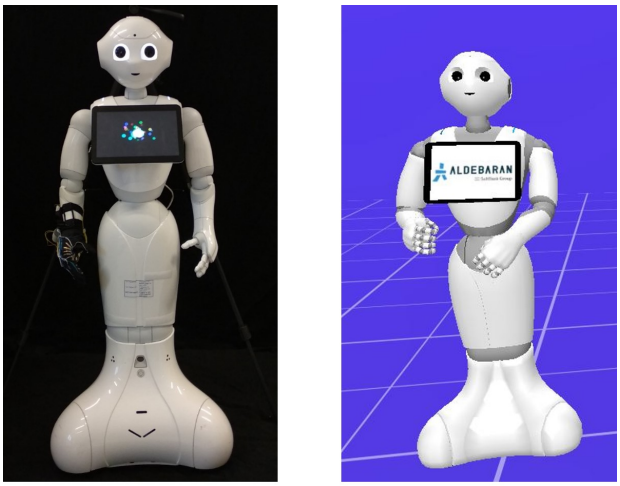


Figure 2.17: The real Pepper robot in our lab (Left) and virtual Pepper robot in *Choregraphe* (Right).

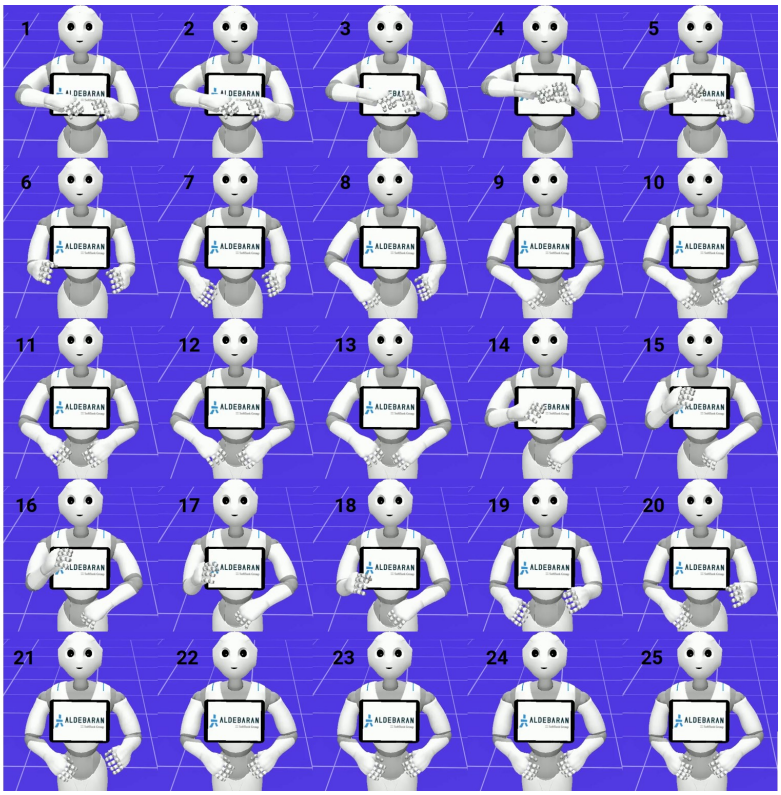


Figure 2.18: Generated gestures on the Pepper robot. You can see the video from Link.

retargeting operation on the gesture with positions, the joint rotation angles are obtained, which can be applied to the Pepper robot. We generated samples with trained models of different epochs and then transferred to the robot joint motor angles as shown in Figure 2.19 and Figure 2.20, in which green curves, red curves, and yellow curves are relative to the ground truth, the generated gesture with noise 1 and the generated gesture with noise 2, respectively. There are four models used for discussion: the 400th epoch model, the 1060th epoch model, the 2000 epochs model, and the 4080 epoch model. In the GAN training process as shown in Figure 2.16, we can see:

- The model mainly focuses on reducing $L1loss$ and GAN losses fluctuate up and down near the 400th epoch.
- The model near the 1060 epoch starts convergence based on the GAN loss curves (generator loss and discriminator loss) where GAN losses do not change with time.
- The model has finished the GAN loss training and focuses on the frame-wise $L1loss$ near the 2000th epoch and 4080th epoch.

Hence, we would like to show what happen in these training phases. From Figure 2.19, we can see that the generated gestures have learned the main gesture curve trend even though the model is only trained for 400 epochs. For the 400th epoch gestures, the noise play an important role in gesture generation (speech audio and noise guide the gesture generation together). It can be noticed that the two noises are relative to two curves with very different broad trends. In this phase, the model mostly work on the frame-wise loss, namely $L1$ loss. However, the alignment between the speech and audio is not so rigid compared to other kinds of mapping task. For the results in the 1060th epoch, it has learned the main trend and the gestures of two noises have similar main curve trend.

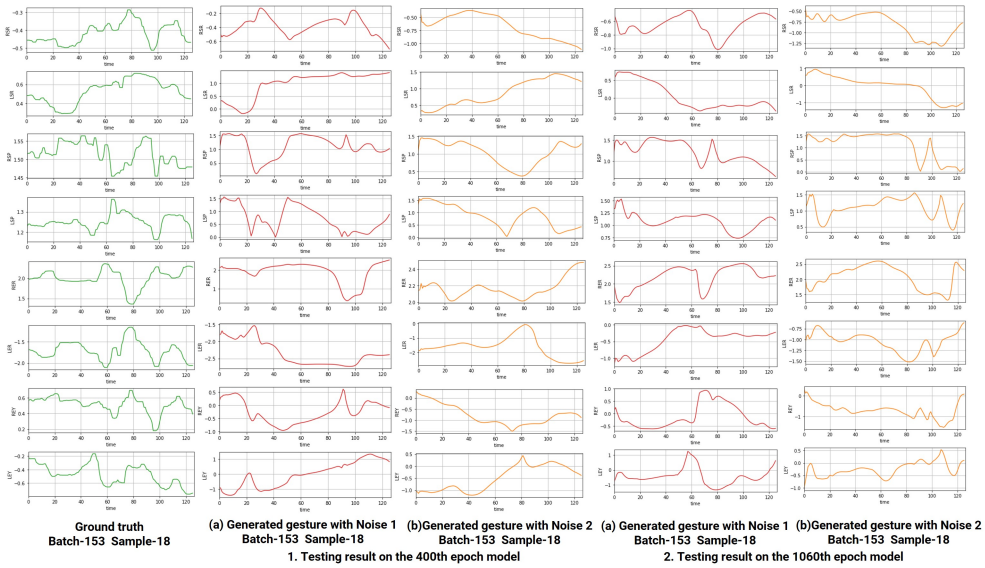


Figure 2.19: The samples of the generated speech-driven gestures and the ground truth. There are two models used here. (1) 400th epoch model (2) 1060th epoch model.

During GAN training, if the generator model has converged, it means that the discriminator cannot classify the generated gesture as a fake sample well. Then, during training, the discriminator feedback gets less meaningful over time. If the GAN continues training past the time when the discriminator is providing quite random feedback, then the generator begins to train with junk feedback, and its quality may collapse. To some degree, the models experience the modal collapse in the 2000th epoch and 4080th epoch when the models start to generate the same pose at the later time steps as shown in Figure 2.20. Hence, the model of 1060th epoch is a potential model for the future real-world human-robot interaction with speech and gesture. Its results have two key advancements: (1) The generated gestures can learn the whole gesture pattern or trend well in general. (2) Although the input is the same one speech audio, *S2GGAN* can generate two different gesture sequences with two different noises and the two gesture sequences are similar in general but with the random variation to some degree, which can lead to a natural lifelong human-robot interaction instead of a repetitive and boring one.

In addition, because the random noise is added to the initial hidden state $\widehat{h_0}$ in the decoder of generator, the joint rotation angles of generated gestures in the first steps are not so stable. However, the situation will be improved with the time steps increasing as shown in LSP (left shoulder pitch) and RSP (right shoulder pitch) sequence of Figure 2.19 2 (a) (1060th epoch,noise 1). The generated gesture data is more smooth because GAN model mostly focus on whether the speech audio and the gesture match with each other. Namely, GAN model pays more attention to the global alignment instead of the frame-level alignment compared with the autoencoder model.

Quantitative evaluation

It is a big challenge to assess the crossmodal mapping with weak or non-strict alignment, just like the speech-to-gesture mapping. Even so, there are some trials on it. In this thesis, we also estimate the generated pose using an Average Position Error (APE) [168] as shown in Equation 2.23, where T is the time steps and is equal to 126; M is the number of testing samples and is equal to 960 (30 batches with batch size 32); $xyz_{\text{real}}(m,t)$ and $xyz_{\text{generated}}(m,t)$

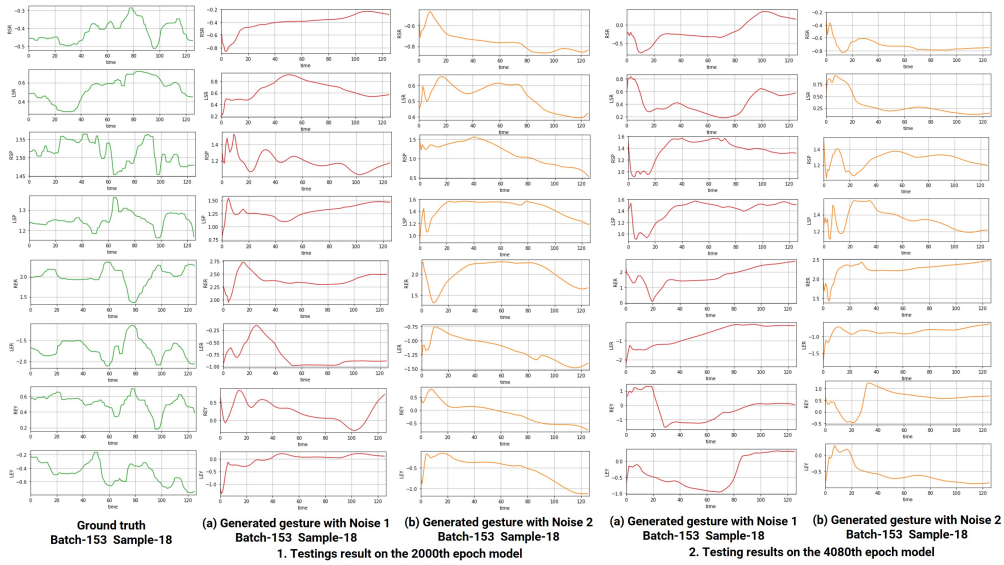


Figure 2.20: The samples of the generated speech-driven gestures and the ground truth. There are two models used here. (1) 2000th epoch model (2) 4080th epoch model

Table 2.1: APE with noise 1 and noise 2

APE(cm)	Noise 1	Noise 2
Head (x, y, z)	7.70, 10.79, 14.78	7.86, 10.58, 14.66
Left shoulder (x , y, z)	4.73, 5.42, 10.76	4.72, 5.25, 10.70
Left elbow (x, y, z)	17.27, 11.59, 20.98	18.85, 11.56, 21.31
Left wrist (x, y, z)	27.02, 23.83, 23.54	27.66, 24.60, 23.65
Right shoulder (x, y, z)	4.73, 5.42, 10.76	4.72, 5.25, 10.70
Right elbow (x, y, z)	20.67, 11.53, 19.77	20.52, 11.07, 19.91
Right Wrist (x, y, z)	23.78, 24.94, 19.71	24.10, 23.89, 20.06

are the the ground truth and prediction of joint position x, y or z of sample m at time step t, respectively.

$$APE = \frac{1}{M \times T} \sum_{m=1}^{M} \sum_{t=T}^{T} \left| xyz_{\text{real}}(m,t) - xyz_{\text{generated}}(m,t) \right| \tag{2.23}$$

Here, we used the 1060 epoch trained GAN model for speech-driven gesture generation. The related APE results of 7 joints are shown in Table 2.1. Firstly, you can find that the generation with noise 1 and the generation noise 2 had similar results, which certify that the random noise can make the generated gestures have a random variation to a certain extent. Then, the head APE and the shoulder APE are small while the elbow and wrist APE are large. Because the elbow joint and wrist joint at the end of the arm have a large movement space and other joints movements have a limited space in real contexts.

It does not mean better that the generated gestures have a lower APE. After all, APE is a kind of frame-wise error but the speech-to-gesture task is an alignment with weak mapping where one speech can map more than one natural gestures. In addition, the object of the work is to generate one-to-many co-speech gestures for a long-term HRI, which will solves the problem that the frame-wise gesture generation model always generate the same and boring gesture mode for each speech. All in all, in the future, the user experiments with

generated robot gesture should be conducted to assess the effectiveness of the one-to-many co-speech robot gesture generation system during human-robot interaction.

2.6 Summary

When the traditional speech-driven robot gesture generation focused on the rule-base method and one-to-one alignment, we built a GAN-based generative model. Compared to the model based on the frame-wise loss, the generated co-speech gestures with our GAN-based model focuses on the global trend instead of the detail of the gesture sequence to see if the gesture and speech map with each other. The qualitative and quantitative evaluation results showed that our GAN model could generate aligned gestures from speech audio. The results also demonstrated that our one-to-many speaking gesturing generation is an effective architecture where the model taking one speech as input can synthesize multiple mapped gestures as the human does. Instead of the one-to-one mapping to generate repetitive behavior, our one-to-many generative model can obtain a natural human-robot interaction. Our new one-to-many *S2GGAN* model for speech-driven gesture generation is promising to be used on other similar cross-modal mapping tasks with the time series as the input and the output. In addition, the model provided a solution to the hard problem, which is how to add noise to the time series input of GAN model or VAE model. The process of building the speech-driven gesture database with the speech audio and the spontaneous 3D gestures was also presented. *S2GGAN* model was trained and tested to generate various speech-driven gestures. Furthermore, the generated gestures with the joint positions were transformed towards robot gestures with the joint rotation angles, which were applied on the Pepper robot. Lastly the generated gestures were qualitatively and quantitatively evaluated.

During *S2GGAN* testing part, the speech-driven gesture sequences mapping with the speech audio were generated. The synthesized co-speech gesture assessment with a quantitative evaluation is still a big challenge because the alignment between the speech and gesture is not so strict. The future work should focus on the real-world human-robot interaction experiments with human in the loop. With the help of the participant validation, we can evaluate the naturalness of the generated gestures.

2.7 Thesis Contributions

The main contributions in this chapter are as follows:

(1) An audio-visual database from the public YouTube TED video collection was built-up. The link to these videos is the following: https://www.youtube.com/user/TEDtalksDirector/videos. The database includes the speech audio data extracted directly from the videos and the associated 3D human pose data extracted from 2D RGB images with the help of *OpenPose* interface and $3D-pose-baseline$ library.

(2) A new temporal GAN framework of the cross-modal generation is proposed for the alignment between the human-like gestures and the speech. Moreover, our model can use one speech audio (with multiple noises) to generate multiple human-like 3D gesture series, which can lead to a natural instead of boring and repeated human-robot interaction. The model solves the problem of how to add the noise for the generative model with time series source as the input because the temporal noise sequence is hard to generate for temporal GAN. In our generator of GAN, the audio speech as a kind of the time sequence input is represented by the encoder part as a vector, which contains all temporal information of the audio speech. Then the decoder of the generator takes the audio representation and the noise directly to generate the synchronous gesture sequence.

(3) The generated gestures given the speech audio are applied to Pepper robot after 3D pose retargeting processing from the generated 3D pose to the Pepper robot pose, which certifies that the generated gesture can be used in real setups.

In total, compared with traditional one-to-one mapping models from time series (for example the speech text and the speech audio) to time series (for example the action and the video), we came up with a new one-to-many solution for speech-driven gesture generation, which is promising. As the noise of the conditional GAN model with time series as input is hard to be added towards the generator part [150], our GAN model can overcame this problem.

This work has been published at [190].

Chapter 3

Speaking robot face action synthesis

3.1 Overview

The natural co-speech facial action as a kind of non-verbal behavior plays an essential role in human communication, which also leads to a natural and friendly human-robot interaction. Compared to other types of non-verbal behaviors, the facial actions are more expressive and effective to show emotion and attention during interpersonal communication and human-robot interaction. However, a lot of previous works for robot speech-based behaviour generation are ruled-based or handcrafted methods, which are time-consuming and with limited synchronization level between the speech and the facial action. Based on the Generative Adversarial Networks (GAN) model, in this chapter, we developed an effective speech-driven facial action synthesizer, i.e., given an acoustic speech, a synchronous and realistic 3D facial action sequence is generated. In addition, a mapping between the 3D human facial actions to a real robot facial actions that regulate Zeno robot facial expressions is also completed.

3.2 Introduction

Recently, non-verbal behavior generation has drawn more and more attention of researchers from many research areas including computing animation and robotics [191] [192] [193]. Non-verbal behaviors including gaze, gestures, and facial actions can assist the verbal expressions to convey clearer meanings in contrast to speech-only communication and intention. It also can help building trust during a real or virtual communication [194]. The co-speech facial action as a non-verbal behavior plays a significant role in human-human communication as they can express rich meanings including the emotion information in the whole facial expression and the verbal content information in the lip or mouth action [132]. In order to make a natural and friendly human-robot interaction, it is necessary to endow a social robot with synchronous and realistic facial actions. However, it is very challenging to generate aligned facial actions mapping with speech in long-term human-robot interaction. Most of the previous researches used the handcrafted or rule-based approaches [39] for offline facial action generation based on speech. These methods are time-consuming and have a limited continuity level of the successive facial actions.

With the development of deep learning technology, more and more generative models for the time series generation were developed, for example, autoencoder model, seq2seq model, the model with the normalizing flows, and GAN (Generative Adversarial Networks) model [195]. Researchers have explored many areas with a generative model for time series generation, for example, music generation [196], human social trajectories generation [197], and gesture generation [191]. These methods also can be used for expressive co-speech facial action generation and simplify the robot facial action generation process, which is a kind of cross-modal mapping task. The trained generative model for facial action synthesis can be used in real-time and long-term human-robot interaction for a social robot.

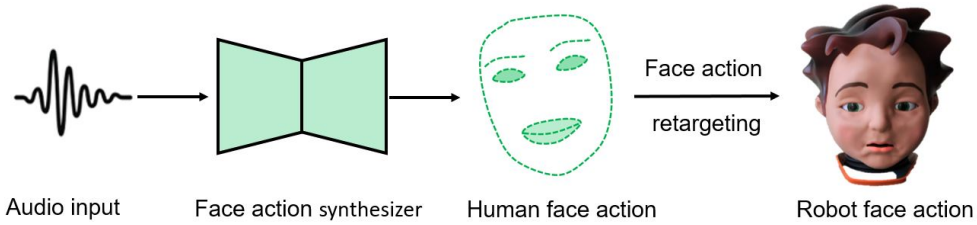


Figure 3.1: The pipeline of robot face action generation from speech. The facial action synthesizer based on the temporal GAN model takes the acoustic speech as input and outputs the aligned human 3D facial actions. The robot facial actions with control signals of robot facial motors are obtained from the human 3D face action during the facial action retargeting part. These motor control signals can be applied to Zeno robot face during human-robot interaction.

In this chapter, we built-up a temporal GAN framework for cross-modal mapping task, which can be applied to generate realistic facial action aligned with the speech audio in a real-time human-robot interaction environment. The basic GAN model is hard to train for cross-modal mapping tasks. Speech-to-facial action generation is not a strict mapping task where humans can conduct many possible natural and simultaneous facial action modes for the same speech, for example, in different emotional states, which makes the GAN more challenging to train towards convergence. To tackle this problem, we built our temporal GAN architecture based on $WGAN-GP$ (Wasserstein Generative Adversarial Networks-Gradient Penalty) [198] model and introduced the L_1 loss in the generator loss function inspired by the $pix2pix$ model [180].

The human face has more than 40 muscles, controlled for facial actions while speaking. However, it is challenging to equip the face of a robot with such a significant number of actuators in order to express rich facial actions. Our research used the Zeno robot, a small humanoid with an expressive face for human-robot interaction, especially used for diagnosis and treatment of autism spectrum disorder (ASD). It has nine motors in its head, two motors for head movement, and seven motors for facial action control. Hence, in this chapter, we completed the facial action retargeting task from 3D human facial landmarks to the robot facial action with related motor control signals. The pipeline of robot facial action generation is as shown in Figure 3.1.

3.3 State of the Art

3.3.1 Generative Model

Generative models, including the model based on Naive Bayes [195], Variational Autoencoder (VAE) [195], Generative Adversarial Networks (GAN) [195], and the model based on the normalizing flows technology [199], have been of high interest to researchers on the image generation tasks and time-series data generation tasks. Habibie et al. [193] proposed a recurrent variational autoencoder model to produce human motions given some control signals, which can be applied for the sequence prediction task. In the paper [180], Isola et al. built an image-to-image translation network based on the conditional GAN model for image generation where the generation loss function also took the L_1 distance into consideration in order to obtain better generation results and to simplify the training process of GAN model. Heter et al. [200] came up with a probabilistic and controllable model for motion synthesis using normalising flows technology. The generative architecture as a probabilistic model

can achieve a one-to-many mapping given multiple control signals, namely style-controllable generation.

3.3.2 Facial Image or Animation Generation

Co-speech video or animation generation with facial action is not a new research topic, which has been explored for decades [201] [202]. The paper [142] built a crossmodal mapping model that inputs the speech text with semantic information and speech audio with rhythmic features and outputs the videos with talking faces. Vougioukas et al. [192] built a temporal GAN model for speech-driven face animation generation. The GAN model used one static image and one speech audio as input and outputted realistic aligned image sequences with the face. In order to improve the randomness of the generated face image sequence, the generative architecture contained a noise generator to produce the noise time series, which was added, respectively to the representation information of each overlapped audio clips in the face generator model of GAN. The generator loss function also considered the L_1 reconstruction loss except for the basic GAN loss, which can improve the generation results. In the paper [203], Zhou et al. built up an LSTM-based expressive face animation generation model with a self-attention encoder. The model took one unseen speech audio and one static speaker image as inputs. With the help of the disentangled learning skills in the model, the model can achieve disentanglement of content and style in audio. The model can generate different talking animations with the same speaker style as the one in the input of the static image. Namely, it is speaker-aware speaking head animation generation.

Both methods above produce the co-speech face or head image sequence. Other researchers focused on the face key points (landmarks) position generation used to control the virtual face avatar in a simulation environment. Sadoughi et al. [204] proposed a conditional sequential GAN (CSG) model to generate the talking lip actions. The model used the spectral and emotional speech features as conditional input of the generator of GAN to synthesize emotion-aware lips action with key point coordinates, which were utilized to regulate the virtual face. Abdelazi et al. [205] described a new co-speech facial movement generation structure that can be exploited for the animated face on smart mobile phones. The model jointly used audiovisual information, including the speech audio and one static face image as input to synthesize the aligned 3D facial action. However, these facial action generation models were only applied in the animation situations and still not explore whether it is effective and whether it can make a difference in the real humanoid robot with face actuated skin.

3.3.3 Robot Facial Action Generation from Speech

Multimodal robot behaviors with speech, co-speech gestures, and facial actions are essential in a natural and friendly human-robot interaction. Particularly, the co-speech facial action generation is an active research area as the facial actions convey more emotional information and speech content information than gestures. Aly et al. [132] built up a multimodal robot behavior synthesis system used on an expressive robot ALICE, in order to imitate natural multimodal human-human interaction. The system can generate speech-related gestures and co-speech facial expressions, which led to an effective narrative human-robot interaction. However, the robot facial action generation method is rule-based. The generated facial action sequences have a limited continuity level in the temporal domain. In the paper [206], the laughter-driven facial motions were generated for a female android robot with face skin. However, the robot facial action generation was rule-based with limited facial action patterns. Therefore, it is challenging to generate facial actions in real-time and long-term human-robot interaction.

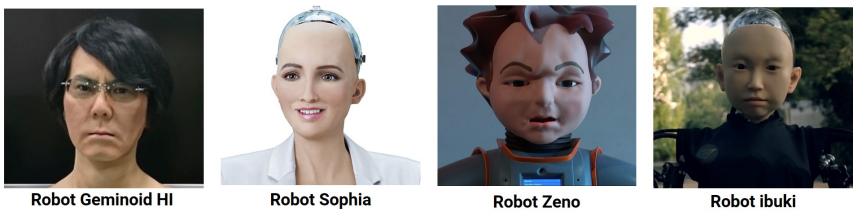


Figure 3.2: The face robot platforms: the robot Geminoid HI, the robot Sophia, the robot ibuki, and the Robot Zeno.

3.3.4 Face robot platform

The robot face with robotic skin plays a vital role in natural human-robot interaction. Humans always convey the emotional state through facial expression during interpersonal communication. There are more and more human-like robots developed with artificial facial skin, such as the robot Geminoid HI, the robot Sophia, the robot ibuki, and the Robot Zeno used in this chapter, as shown in Figure 3.2. These robots with silicon skin on their faces have high-level anthropomorphism and can conduct basic speech-driven face actions or expressions. However, it is still a big challenge to express complex facial action like human beings because the robot has limited face motors to control the facial movement compared with more than 40 face muscles for facial expression. Hence, there are still many works that should be done to improve the face robot platforms.

3.4 Methodology

3.4.1 Problem definition

Speech-driven Facial Action Generation: It is a crossmodal translation task with time series both as input and output. Given one speech audio $S^m = [s_t^m]_{t=1:T}$ as input, the model attempts to produce one 3D facial action sequence $A^m = [A_t^m]_{t=1:T'}$. Namely, the generative model tries to learn a relation function $F_{mapping}$ to maximize the conditional probability $P(A^m|S^m)$ to generate the natural and aligned facial action sequence. Here, T and $T^{'}$ are time steps of the speech audio as input and the facial action sequence as output, respectively, and they are different from each other because the digital speech audio and the facial action sequence have different sampling rates. m in the model means m^{th} mapping task.

$$\mathbf{A^m} = F_{map}\left(\mathbf{S^m}\right) \tag{3.1}$$

Facial Action Retargeting: The problem is to map the human facial actions A^m with the 3D positions of the face key points to the robot facial action sequence $C^m = [c_t^m]_{t=1:T'}$ with the face motors' control signals. The mapping task for face action retargeting consists in getting a function that finds the relation between human facial action and robot facial action. The final function can make the appearance of human facial and the appearance of robot face as similar as possible at each time step.

$$\mathbf{C}^m = F_{retarget}\left(\mathbf{A}^m\right) \tag{3.2}$$

3.4.2 Face action generation from speech

This section describes our novel proposed facial action synthesizer from speech with temporal GAN. The speech-to-face-action GAN ($S2FGAN$) architecture is shown in Figure 3.3. $S2FGAN$ model is made of a generator and a discriminator. The generator with a sequence

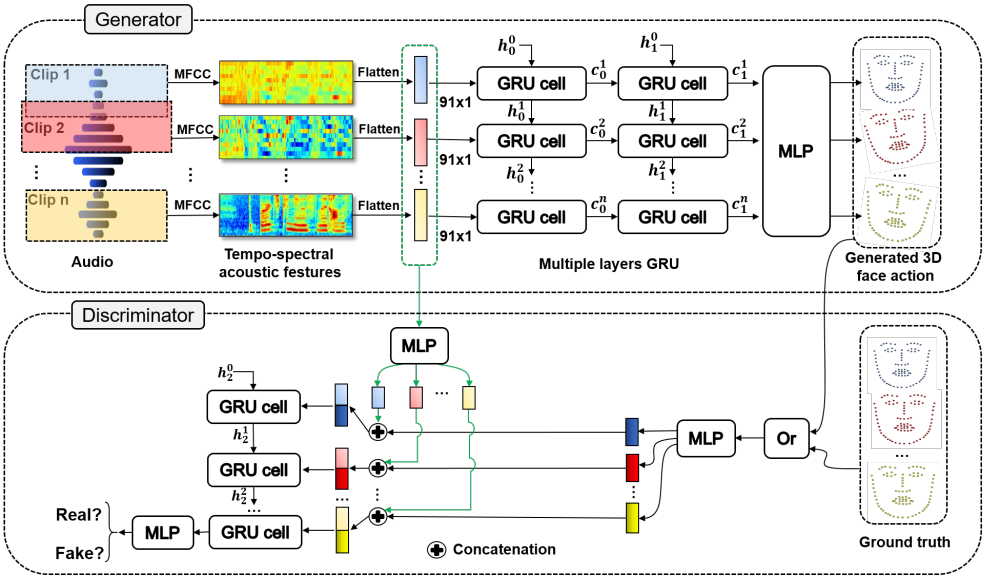


Figure 3.3: The *S2FGAN* architecture. The model has a generator and a discriminator. The generator takes the spectral features of speech as input and outputs the synchronous 3D facial action data. The discriminator with the speech audio and generated/real facial action sequence as inputs try to classify whether the speech and the facial action sequence align in the temporal domain.

model takes the temporal representation of speech audio as input and outputs the mapping gesture. The discriminator is employed to differ whether the speech and the facial action match with each other.

The generator comprises two layers of GRU (Gated Recurrent Unit) and MLP (Multilayer Perceptron) layer. Firstly, Mel-frequency Cepstral Coefficients (MFCCs) as audio representation are extracted from the overlapped audio clips. The MFCC feature sequence is input in the batch normalization layer following two layers GRU of the generator. The following MLP layer takes the latent representation of the former GRU layer to generate the synchronous 3D facial action sequence mapping with the speech audio. Each frame of the facial action sequence contains 3D positions of 68 face landmarks. The discriminator works to distinguish whether the facial action sequence and the speech audio match with each other. The audio clip representations and the facial action sequence are input to two MLP layers and decode to 100-dimensional features and 50-dimensional features each time step, respectively. The following concatenation layer fusions the two modal features in each time step, whose output is input to a GRU layer. In the final GRU cell, an MLP layer is followed to classify whether the speech audio matches the 3D facial action sequence.

The loss function of our *S2FGAN* model comprises two parts, namely L_1 loss part and the standard conditional GAN loss part, actually the Wasserstein loss and the gradient penalty used in WGAN-GP model.

Basic GAN model often experiences the training instability problem. The Wasserstein GAN (WGAN) model makes a more stable training than basic GANs [207]. WGAN can also produce samples with low quality and suffer from convergence problems during the training process. WGAN introduces a weight clipping skill to enforce a Lipschitz constraint on the discriminator (namely, the critic named in WGAN) to address these problems, which also can result in gradient explosion/vanishing without careful tuning of the weight clipping parameter. In WGAN-GP [198], the authors proposed an alternative skill to the weight clipping, namely, adding the gradient penalty to the discriminator loss, which led to a more stable

training process. The WGAN-GP loss contains the generator loss $\mathscr{L}_G$ and the discriminator loss $\mathscr{L}_D$, as shown in Equation 3.3 and Equation 3.4, respectively. Where, S^m is the speech audio input; A^m is the ground truth face actions aligned with the speech S^m; the sample A^n from the sampling uniformly along straight lines between pairs of points sampled from the data distribution of real facial action sequences and the generator distribution.

$$\mathscr{L}_G = -\mathbb{E}_{S^m}[D(S^m, G(S^m))] \tag{3.3}$$

$$\begin{aligned}\mathscr{L}_D =& \mathbb{E}_{S^m}[D(S^m, G(S^m))] - \mathbb{E}_{S^m,A^m}[D(S^m, A^m)] + \\ & \lambda \mathbb{E}_{A^n}[(\|\nabla_{A^n} D(A^n)\|_2 - 1)^2]\end{aligned} \tag{3.4}$$

Inspired by the *pix2pix* model [180], we introduced a L_1 reconstruction loss to improve the realistic co-speech facial action generation. The L_1 loss is pixel-wise in the image translation task with *pix2pix* model, while we used the frame-wise L_1 loss for the facial action sequence, as shown in the Equation 3.5. The final discriminator loss keeps same and the L_1 loss is added to the generator loss to get the final generator loss $\mathscr{L}_{G-all}$ as shown in Equation 3.6. Where, λ is an empirical hyperparameter during *S2FGAN* model training, which is to balance how much contribution $\mathscr{L}_{L1}$ or $\mathscr{L}_G$ make for all the loss.

$$\mathscr{L}_{L1}(G) = \mathbb{E}_{S^m,A^m}[\|A^m - G(S^m)\|_1] \tag{3.5}$$

$$\mathscr{L}_{G-all} = \mathscr{L}_G + \lambda \mathscr{L}_{L1}(G) \tag{3.6}$$

3.4.3 Robot face action mapping

Some previous works try to use a few face motors with limited degrees of freedom to code face actions or expressions on the robot face [208]. Our facial action retargeting task consists in mapping human facial action to robot facial action. The mapping objective is to approximate the human face appearance with a limited number of robot face actuators. In this chapter, we use a Zeno robot to present the synchronous generated facial action sequence with the speech audio. There are nine motors used for the robot head, two motors controlling the robot head pitch and head yaw, two motors controlling the left eye and the right eye turns and five motors for skin-based face appearance regulation. Our research focuses on facial expression, so Zeno's five face motors are used during the retargeting task. The five motors and related functions for facial expression are as follows:

- The frown motor: to control the robot facial skin areas of the left eyebrow, the right eyebrow, and the forehead for the frown motion.
- The open eye motor: to regulate the robot left and right eyelids open and close for the eye blink action.
- The motors of left and right mouth corners: to control the smile motion of the left and right mouth corners.
- The mouth motor: to adjust the mouth opening and closing actions.

Each motor's control signal of Zeno robot is a continuous value ranging from 0 to 1. The retargeting process from human facial action to robot facial action is as shown in Figure 3.4. The human facial action includes 3D positions of 68 landmarks, and five motors of Zeno with skin regulate the robot's facial expression.

We name the distance between the 38^{th} landmark and the 42^{nd} landmark as h_{1r}, the 39^{th} landmark and the 41^{st} landmark as h_{2r}. The right eye wide y_{kr} is the distance between the 37^{th}

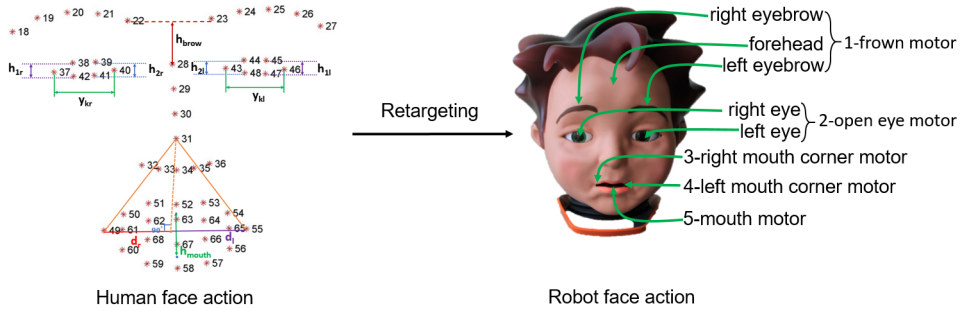


Figure 3.4: Facial action retargeting overview The human face contains 68 human landmarks in each frame. The robot face has five motors controlling the eye, the forehead, the mouth, and left/right mouth corner for smile

landmark and the 40th landmark, which is used to normalize the the open degree of the eye as different persons have the different eye sizes. Apply the same rule for the left eye to get the h_{1l}, h_{2l} and y_{kl}. Because Zeno has only one motor to control two eyelids, we calculate the average for the two eyes. Then, we can get the scale for eyelid motor as shown in Equation 3.7.

$$S_{eye} = \frac{\frac{h_{1l}+h_{2l}}{2y_{kl}} + \frac{h_1r+h_{2r}}{2y_{kr}}}{2} \tag{3.7}$$

To obtain the eyebrows motor scale, we need to calculate the distance between the midpoint of the 22nd landmark and the 23rd landmark and the 28th landmark, h_{brow}. So, the scale for eyebrows as shown in Equation 3.8. Here, we divide $(y_{kl}+y_{kr})/2$ is to reduce the influence of different face sizes of people on the results of the mapping task from 3D facial action to robot motor action.

$$S_{\text{brow}} = \frac{h_{\text{brow}}}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{3.8}$$

The scale for mouth motor can be obtained as shown in Equation 3.9. Here, h_{mouth} is the distance between the midpoint of the 52nd landmark and the 63rd landmark and the midpoint of the 67th landmark and the 58th landmark.

$$S_{\text{mouth}} = \frac{h_{\text{mouth}}}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{3.9}$$

The scales of two motors controlling the left and right corners of the mouth can be obtain from Equation 3.10 and Equation 3.11. The d_l is the distance between the landmark55th and the foot of the perpendicular through the landmark 31st. Similarly, the d_r is the distance between the landmark49th and the foot of the perpendicular through the landmark 31st. And d_l and d_r can be calculated based on the law of cosines.

$$S_{smile_l} = \frac{d_l}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{3.10}$$

$$S_{smile_r} = \frac{d_r}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{3.11}$$

Since robot motor control signals in Zeno system range from 0 to 1, normalization operation for the scales should be done as shown in Equation 3.12. That is to say, find the maximum and minimum of every scale which are applied to get the final control signal of face motors. Where, $s \in \{S_{eye}, S_{\text{brow}}, S_{\text{mouth}}, S_{smile_r}, S_{smile_l}\}$. When selecting the maximum and the minimum, we also remove some noise data, such as few strange points that are very

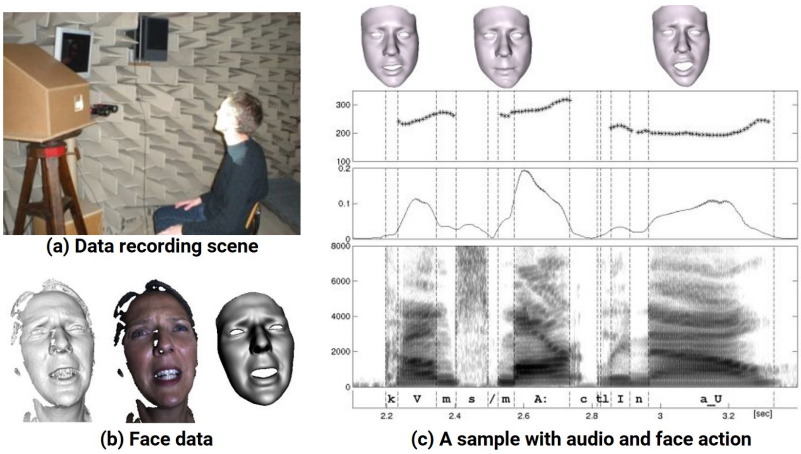


Figure 3.5: Biwi 3D Audiovisual Corpus of Affective Communication dataset [209] (a) the data recording scene: one speaker sits in front of the 3-D scanner in the anechoic room while watching one of the eliciting videos clips.(b) the face data including face 3D reconstruction, the corresponding colorful texture mapped on 3D reconstruction, and the personalized face template deformed to fit the specific frame. (c) a sample with 3D face and speech.

big or very small and far away from the data distribution. If we do not remove those noise points, we will get the wrong maximum and minimum values from the data samples, which will lead to the mapped robot motor signal far smaller or bigger than the real one.

$$\text{norm}(s) = \frac{s - s_{\min}}{s_{\max} - s_{\min}} \tag{3.12}$$

3.5 Database and preprocessing

3.5.1 Database

In this chapter, we used the open database-Biwi 3D Audiovisual Corpus of Affective Communication dataset [209], which was developed at ETH Zurich. The corpus contains 1109 sentences uttered by 14 native English speakers, including six males and eight females, aged between 21 and 53 (average age of 33.5). A real-time 3D scanner and a professional microphone were utilized to obtain the speakers' facial action and synchronous speech audio during the data recording process. The data recording scene is shown in Figure 3.5(a) The dense dynamic face scans were obtained with a sampling rate of 25 frames per second. The related face data can be seen in Figure 3.5(b) and the 3D face data include the basic face 3D reconstruction, the corresponding colorful texture mapped on 3D reconstruction, and the dense 3D face. Moreover, the RMS error in the 3D reconstruction is about 0.5 mm, which is good enough for our facial action generation task. For the dataset development, the participants imitate the forty short English sentences extracted from film clips. For each sentence, the subject should speak two times, one with a neutral state and one with an emotional state, which is the same as the film clip's emotion. One sample with speech phonological representation and 3D face is as shown in Figure 3.5(c). In this chapter, the speech audio contains intrinsically emotional information , so we did not take the emotion label into consideration for co-speech facial action generation task. Namely, the work in this part only models the crossmodal mapping between the prosody-based speech and the face action.

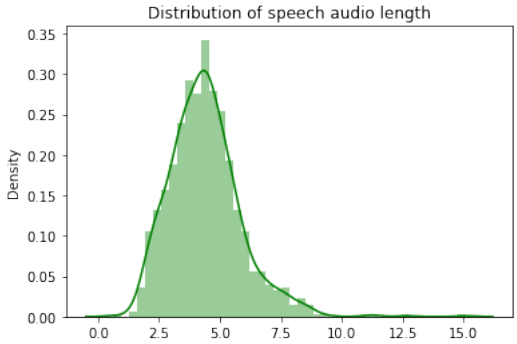


Figure 3.6: Distribution of speech audio length.

3.5.2 Pre-processing

The pre-processing step includes speech audio spectral feature extraction with MFCC [210], 3D face landmarks extraction from face images in the database library, and how to align the speech and facial action in the temporal domain and so on. The database is small and not enough to train the 1D CNN with massive parameters. MFCC is a rule-based method and has a good performance on prosody-based phonological feature extraction. Hence, instead of 1D CNN for speech audio representation, the MFCC feature is used. The 3D face action is based on the RGB face image and the relative texture image. Firstly the 2D face landmarks are detected from the RGB image. Then 3D face action can be extracted from the 2D face action with the help of the texture information. This thesis does not use the 2D face action directly because the 3D face action leads to a more precise robot face controlling in the later robot face mapping task. Then, the necessary face key points were selected that will be used in robot face retargeting task. Lastly, the median filtering and standardization processing were applied on the 3D face action.

Alignment between Speech and Facial Action Sequence

In this chapter, we used the same time size for each speech audio to simply the training process. From the distribution of audio length as shown in Figure 3.6, we know that most audios are longer than 2.5 seconds. There are 1096 files in our database, of which 1095 are longer than 1 second, 1072 are longer than 2 seconds, 926 are longer than 3 seconds, 645 are longer than 4 seconds. The the speech audios of short duration cannot contain the enough whole prosody-based features. Hence, we chose 3 seconds as the time size for *S2FGAN* training to use as many samples as possible. The audios longer than 3 seconds were cut into 3 seconds, and the samples with audio size shorter than 3 seconds were deleted from the database. Meanwhile, the sampling rate of the face image is 25 fps. In addition, we also deleted some samples whose audio size was more than 3 seconds but face images less than 3 seconds. Finally, 788 samples are got from the original dataset for *S2FGAN* training.

Because the speech audio and facial action series have different sampling rates, namely 44100 Hz for audio and 25 fps for facial action, the whole speech audio was divided into audio clips to align the facial action and audio in the temporal domain. Namely, one frame corresponds to 1764 audio frames. Considering the facial action time series's temporal dependence, we used the overlapped audio clips with 3528 audio frames centered on the related facial action frame, and the stride of the overlapped audio clips was 1764.

Speech Audio Feature Extraction

MFCC (Mel-Frequency Cepstral Coefficients) is often used for acoustic speech representation in speech recognition and other related speech audio tasks. MFCC feature of speech audio is the one in the frequency domain using the Mel scale based on the human ear scale. MFCC, as frequency domain features, has been certified as an effective speech representation in the past research [211]. So, we used the MFCC as the overlapped audio clips from the whole speech audio in this chapter. The related MFCC processing pipeline is shown in Figure 3.7. The process contains four steps:

- Pre-emphasis, frame blocking & widowing: Pre-emphasis processing is a filter that emphasizes the higher frequencies. The transfer function is shown in Equation 3.13, where a is used to control the slope of the filter. The frame blocking process is to cut the audio into short clips across which the speech signal is assumed to be stationary. The windowing processing can use the Hanning or Hamming windows [212] to enhance the harmonics, smooth the edges, and reduce the edge effect while taking the FFT on the signal.

$$H(z) = 1 - a \cdot z^{-1} \tag{3.13}$$

- Fast Fourier transfer (FFT): Each window frames (audio clip) was transferred to the magnitude spectrum.

- Mel-filter bank: The magnitude spectrum above is fed to a set of band-pass filters known as a mel-filter bank. A *mel* refers to a frequency that human ears can perceive to physical frequency because the human auditory system does not sense the physical frequency of the tone linearly. The approximate mel can be obtained from the physical frequency f as shown in Equation 3.14.

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.14}$$

- Discrete Cosine Transform (DCT): DCT is used to extract the MFCC features.

The MFCC processing used 25 milliseconds as the length of the analysis window, 10 milliseconds as the step between successive windows, 26 as the number of filters in the filter bank, 13 as the number of cepstrum, and 512 as fast Fourier transform (FFT) size. The input audio sequence of MFCC is a 1D series with a length of 3528. Each audio clip corresponding to one face frame extracted an MFCC feature with size $7 * 13$ as shown in Figure 3.8.

3D Face Landmarks Detection

To get the 3D face landmarks, we first got the 2D face landmarks from the 2D face image frame by frame based on the Dlib library. Dlib is an open programming toolkit with machine learning algorithms and tools for building up complex software to resolve real-world problems. It has been used widely in industry and academic research. Here we used the Dlib image processing interface [213] to detect 68 2D face landmarks. The pre-trained face landmark detector inside the Dlib library can extract the location of the 68 face landmarks (x, y)-coordinates that map to face structures on the face, as shown in Figure 3.9. The clear indexes of the 68 coordinates can be seen in Figure 3.4.

In our case, the triangle mesh texture recorded in the database is the corresponding RGB file. Firstly, we have detected 2D face landmarks located in the RGB face image. The detected landmark location is the same as the landmark location in the texture image. The relation between texture image and 3D mesh can be learned from the depth image. Then

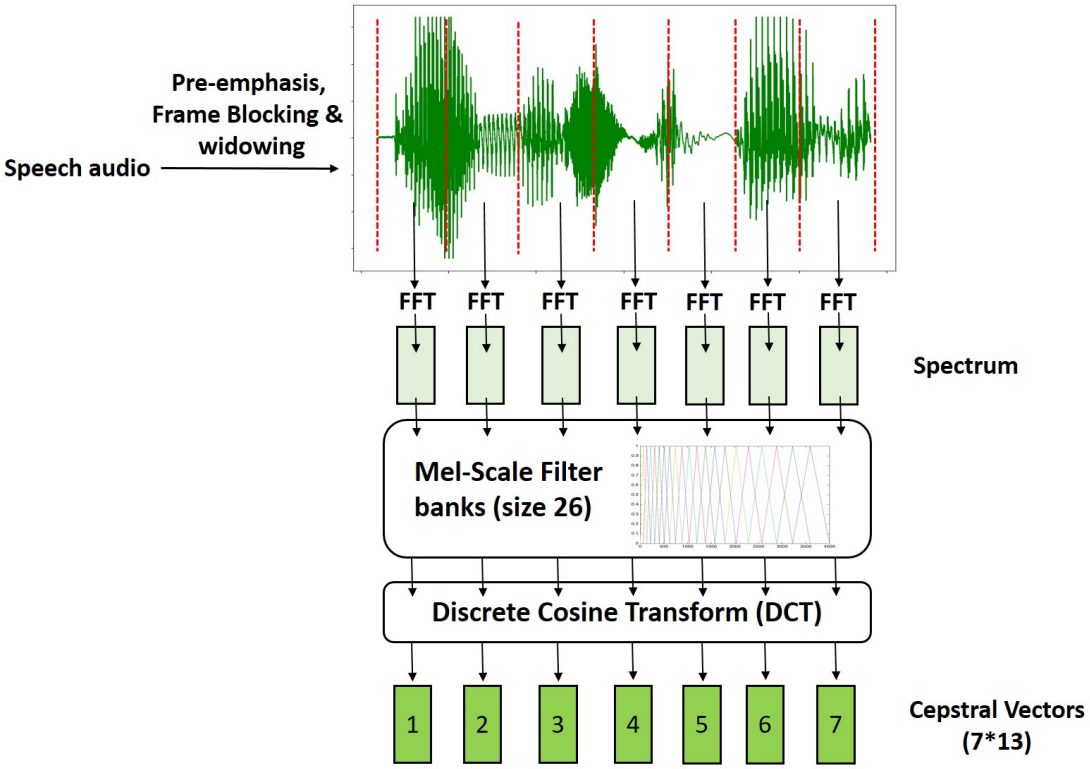


Figure 3.7: MFCC processing pipeline. Four steps: (1) Pre-emphasis, frame blocking & widowing. (2) Fast Fourier transfer (FFT). (3) Mel-filter bank. (4) Discrete Cosine Transform (DCT).

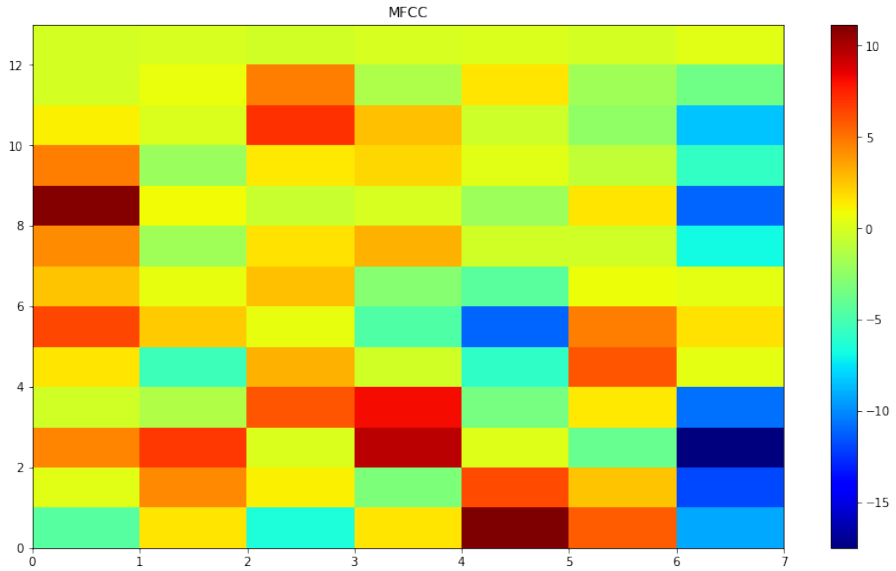


Figure 3.8: The MFCC feature of one speech audio clip with a length of 3528. Then each MFCC feature extracted from each audio clip will be flatten as the input of each GRU cell in Generator.

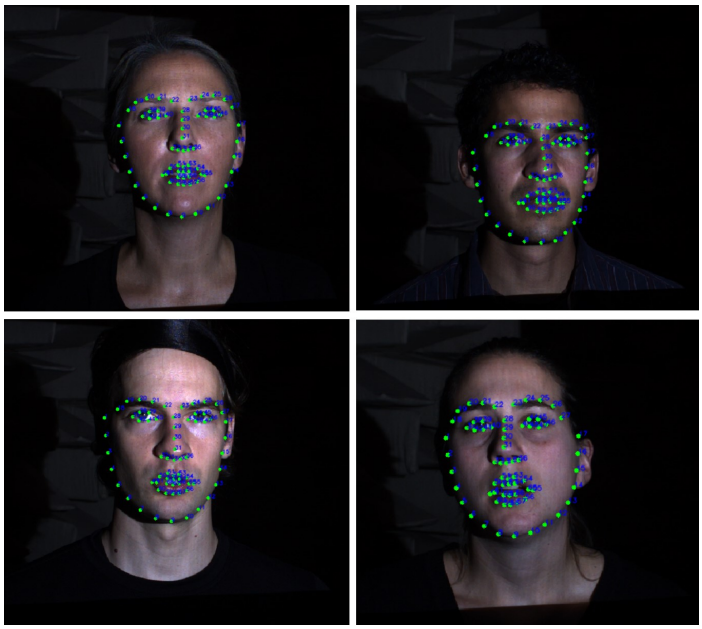
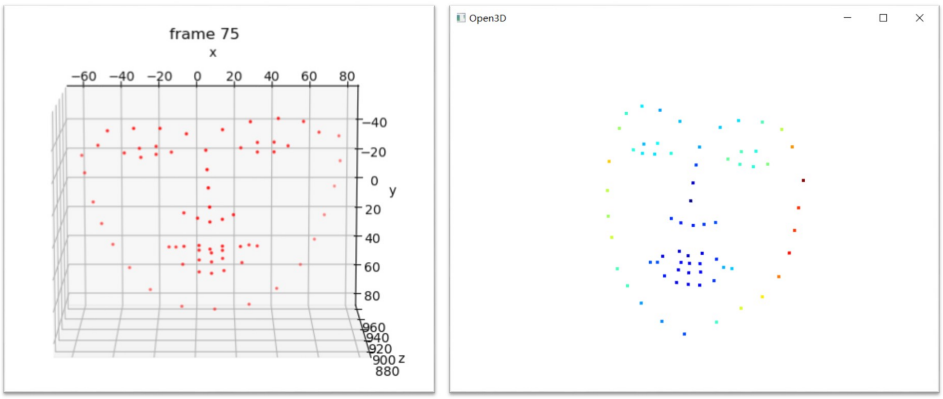


Figure 3.9: The face key point detection results of 4 examples with Dlib face detection interface. The interface can detect 68 key points from the face image.

from the 2D position, we can directly extract the 3D positions of the 68 landmarks. As shown in Figure 3.10(a), the 68 3D facial key points are plotted in 3D coordinate system. The face looks weird because of the 3D object projection. We also use the 3D Open3D [214] to visualize the 3D face key points clearly. Then the extracted 3D face action sequence is proceeded with the median filter (window size: 5) to remove the noise introduced during 2D face action extraction 3.11. The final goal of the task is to generate the speaking robot face action, which can be used to control the robot's facial expression. On the Zeno robot, five motors can be used for robot face expression control. As discussed in the robot face retargeting part, 22 3D face key points were used for five robot face motors' control. The redundant 3D face action will mislead the generative model towards convergence and leads to unnecessary computing consumption. Hence, during database building for generative model training, we only selected the 3D position data of those 22 key points instead of 68 ones. Finally, the standardization processing was applied to re-scale features of 3D face action,



(a) 3D key points in 3D coordinate system (b) 3D key points visualization through "Open3D"

Figure 3.10: The 68 3D face key points extracted from 2D key points.

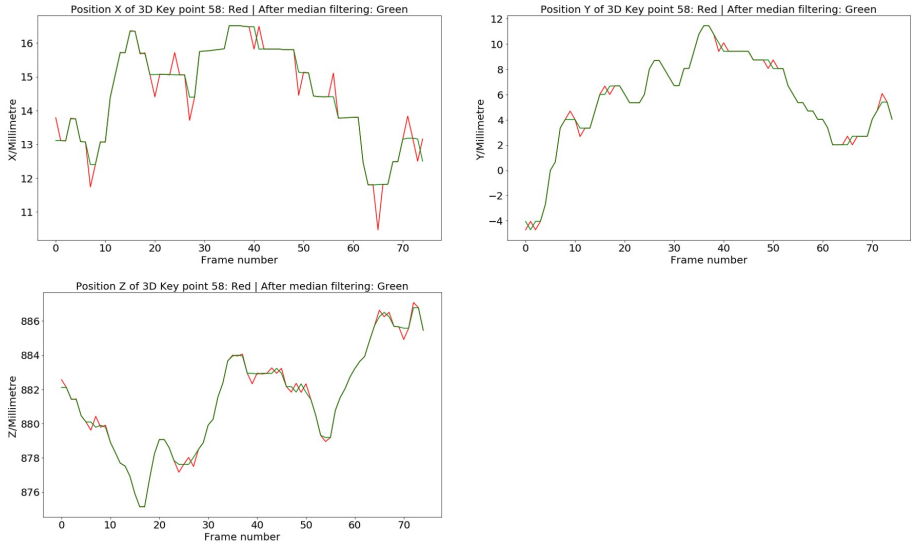


Figure 3.11: The median filter for 3D positions.

facilitating a good and fast training process of the generative model.

3.6 Experiments and results

3.6.1 Model training

The conditional GAN *pix*2*pix* model with L_1 loss explored multiple cross-modal translation tasks with the small dataset with 400 images or less, and it got the receivable testing results finally [180]. Like the *pix*2*pix* model, our speech-face database for *S*2*FGAN* training contains 788 samples (600 samples for training, 90 samples for validation, and 98 for testing) with the speech audios and the 3D facial action sequences. During the training, The batch size is 30, and the time steps of 3D facial action are 75 as the audio time size was fixed to 3 seconds during the *S*2*FGAN* training. The Standardization operation was employed on the 3D facial action data before inputting the *S*2*FGAN* model, and batch normalization procedure was applied in the generator of *S*2*FGAN*, which both can effectively reduce the overfitting problem during model training based on the tuning experiments. We did not use the batch normalization layer in discriminator because the layer can lead to a convergence failure during WGAN training [198]. Both generator and discriminator of *S*2*FGAN* model used the Adam algorithm [189] for optimization during training with the learning rate = 0.0002, the parameter $\beta_1 = 0.5$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-7}$. Moreover, the dropout setting of GRU is 0.1. The number of discriminator iterations per generator iteration is five during training. The model is developed with Tensorflow 2.3, and the training with 10000 epochs was done on a NVIDIA GeForce RTX 2080 Ti GPU for about 35 hours. In different training epochs, the trained models can generate the 3D face action with various performances, as shown in Figure 3.12. In the figure, we used the mouth motor signal controlling lip motion to check the models' ability because the lip motion is most relevant to a speech during speaking than other facial unit motions. We can see that the model performance becomes better with the epochs increasing. The 1000th model still cannot learn the trend of the curve. The 4000th model has learned the main pattern of face action. Furthermore, the 8000th model starts to know the details of the curve. The 10000th model can get the face action trend in general and has learned a more detailed mode. Hence, we used the 10000th epoch model as the final one to explore the effectiveness of speech-driven robot face action through the user assessment experiments.

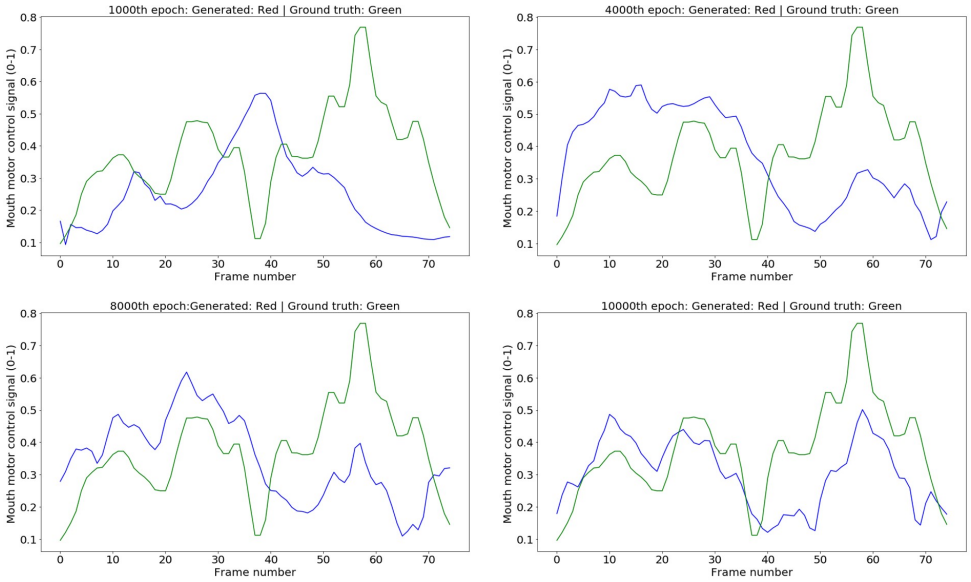


Figure 3.12: The performance of trained generative models in four different epochs. The four epochs are 1000, 4000, 8000, and 10000. Green: ground truth; Blue: generated.

3.6.2 Qualitative evaluation

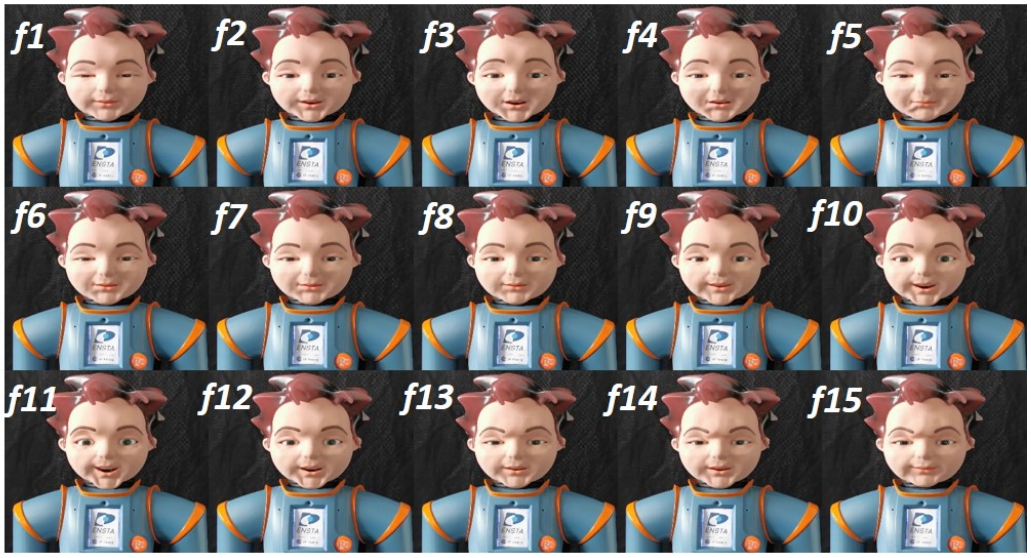
During the testing part, the speech-driven 3D facial action sequences were generated using the trained *S2FGAN* model. Then, the generated facial actions were transferred to the control signals of the robot face motors. The following median filtering is used to smooth the robot motor signal, which can protect the robot motors from the harm of a sudden big rotational speed. Then the filtered robot motor control signals finally were presented on the Zeno robot facial actions with the aligned speech audios.

Applying the generated co-speech robot facial action to Zeno robot, we recorded some videos with the speech audio and the synchronous facial action, and some frames in a generated co-speech robot facial action sequence are as shown in Figure 3.13. The related speech text is $Oh, Mr. Bennet! How can you tease me so?$. The sampling frame is 5 Hz. Each sample of three seconds has 15 frames. The number label is the figure is the frame number. For example, the $f1$ indicates the first frame.

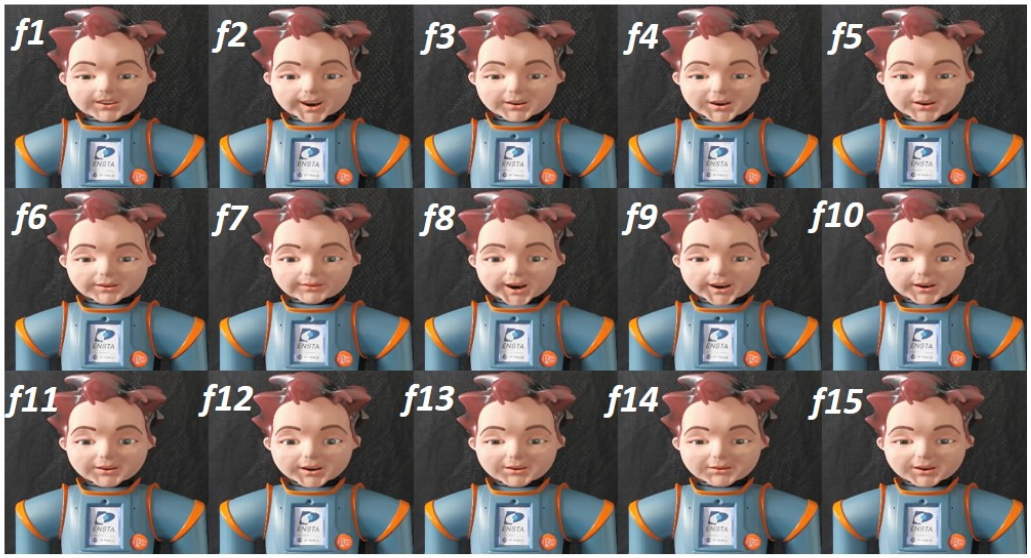
From Figure 3.13, we can see that the Zeno face driven by the generated robot facial action has noticeable movement in the mouth area and the eye area. The forehead area has limited change as the human forehead's noticeable movement often happens in intensely emotional expression instead of the common human co-speech facial actions. Besides, the subjects' forehead mostly remains still during the speech as present in the database.

3.6.3 Quantitative evaluation

The quantitative evaluation of speech-driven facial action or gesture is challenging [168] as the mapping between speech and facial action (or gesture) sequence is a weaker correlation than the image-to-image translation in *pix2pix*, which is a rigid one-to-one mapping. And the lip motion has a stronger relationship with the prosody-based features of the human speech utterance than the actions of other facial units, such as the eye blink, the frown, and smile. In this chapter, we explored the quantitative evaluation for the generated speech-driven facial action sequence with an Average Position Error (APE) [168] as shown in Equation 3.15, where T is the time steps of the robot facial action, equal to 75; S is the number of testing



(a) Generated robot face action



(b) Ground truth robot face action

Figure 3.13: The generated facial action and ground truth on the robot Zeno in one example. The speech text of this example is "Oh, Mr. Bennet! How can you tease me so?". Fifteen frames were sampled from each face action series of three seconds. The sampling frame is 5 Hz. The number in the figure is to show the number of the sampled frames. The video of the ground truth can be seen from the Link and the generated face action video can be accessed from the Link.

Table 3.1: The APE of five face motors in four different epochs

APE (0-1)	Eye	Frown	Mouth	Left mouth corner	Right mouth corner
Epoch 1000	0.504	0.260	**0.327**	0.173	0.174
Epoch 4000	0.494	0.187	**0.223**	0.185	0.172
Epoch 8000	0.397	0.183	**0.221**	0.162	0.238
Epoch 10000	0.409	0.190	**0.187**	0.179	0.199

samples, equal to 98; $f_{real}(s,t)$ and $f_{generated}(s,t)$ are the real robot facial motor control signal action and the generated one of sample s at time step t, respectively.

$$APE = \frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} \left| f_{\text{real}}(s,t) - f_{\text{generated}}(s,t) \right| \tag{3.15}$$

The APE validation results are shown in Table 3.1. In any epoch, the eye motor has the biggest APE as the degree of the eye opening has a weak correlation with speech and the action is mostly random in human speech. We have known that the lip motion relative to the robot mouth motor has a strong correlation with speech. And from Table 3.1, we can see that the APE of the mouth opening motor becomes smaller with training epochs increasing. The result looks like it still has some space to improve. However, it is receivable to model the robot speaking behavior with a limited database for this kind of one-to-many mapping task with weak correlation where different aligned face actions from different speakers can map similar speech audio. And APE is a good assessment method for pixel-wise or frame-wise mapping tasks. It is still a challenge to work well on the speaking robot face action synthesis task. Hence, we will validate the generated robot facial action with user experiments in a later discussion.

3.6.4 Human evaluation

Human evaluation is an appropriate way to assess the speaking face action generation. Hence, we made many videos with robot face action and spontaneous speech audio and then asked participants to watch the videos to validate whether the generated robot facial action is synchronous with the speech, if the facial action is friendly, if the face action is human-like, and whether they can recognize the generated face action and the ground truth.

At first, we randomly selected 15 samples from 98 generated speaking robot face sequences. Thereafter, we used these samples and the related ground truth to make 30 videos with robot face action and spontaneous speech audio for each instance. The videos were uploaded on YouTube in a random order, where we cannot know if the video is with a generated face or a ground truth. Finally, the questionnaire was made through Google Forms as an online survey, as shown in Figure 3.14. That allowed participants worldwide to join the experiment. The questionnaire can be accessed from the Link.

The human evaluation experiment contains two parts: the personality assessment and robot face action assessment parts. The personality assessment is based on the Big Five personality traits [215]. The Big Five personality questionnaire used in this chapter contains 45 questions to assess five subject personalities i.e. Extroverion, Agreeableness, Conscientiousness, Openness, and Neuroticism. These questions can be seen in the Appendix A-*Chapter 9*. Each question is quantified on a five Likert scale: disagree strongly, disagree a little, neither agree nor disagree, agree a little, and strongly agree. In this chapter, we explored the assessment results of generated robot face actions based on the participants' personality. The participants watched the 30 videos (15 with generated face action and 15 with ground truth)

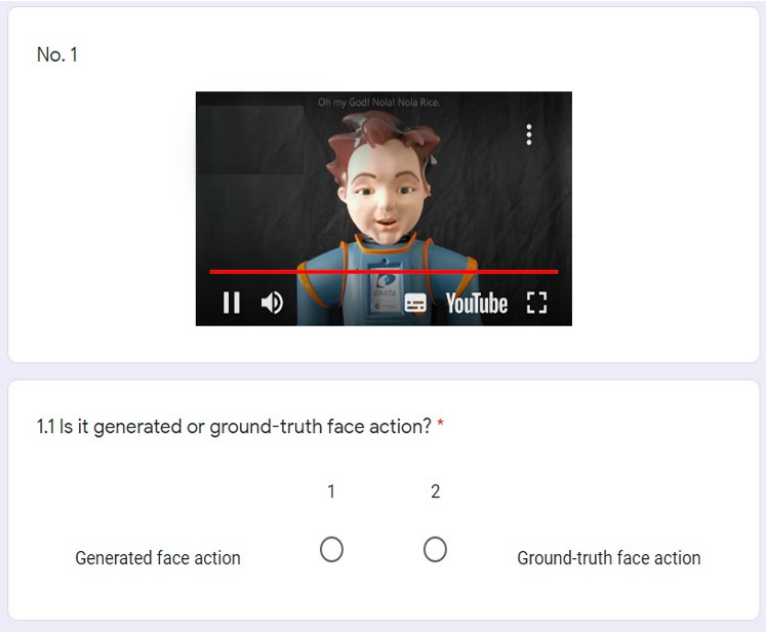


Figure 3.14: The online Google Form questionnaire scene. The details can been seen in the Link

with the robot face action and the spontaneous speech utterance in the second part. Each video refers to four questions shown as following. Each question can be answered with a score selected from a 5-point Likert scale [216] where "1" corresponds to "strongly disagree" and "5" corresponds to "strongly agree".

- Is it generated or ground-truth face action?
- How much does the face action match synchronously with the audio?
- How much is the face action friendly?
- How much is the face action human-like?

All in all, the experiment steps used to assess the synthesized robot action from the speech are as follows:

- The participant access the questionnaire from the Link.
- The participant first answers 45 questions for the personality assessment.
- The participant watches the 30 videos where the robot speaks with related face action and then answers four questions below each video.

In this experiment, 55 subjects participated. All of them are students studying in France, China, Japan, Germany, and the UK. There are 22 females (40%) and 33 males (60%). Due to the Covid19 pandemic, this experiment was only done online through the Google Form questionnaire. No face-to-face experiment could be conducted in this period. Each subject did the experiment once.

Is it generated or ground-truth face action?

In the first question, we asked the participants to recognize the video as a generated speaking face action or ground truth when they did not know the label of the video during experiments.

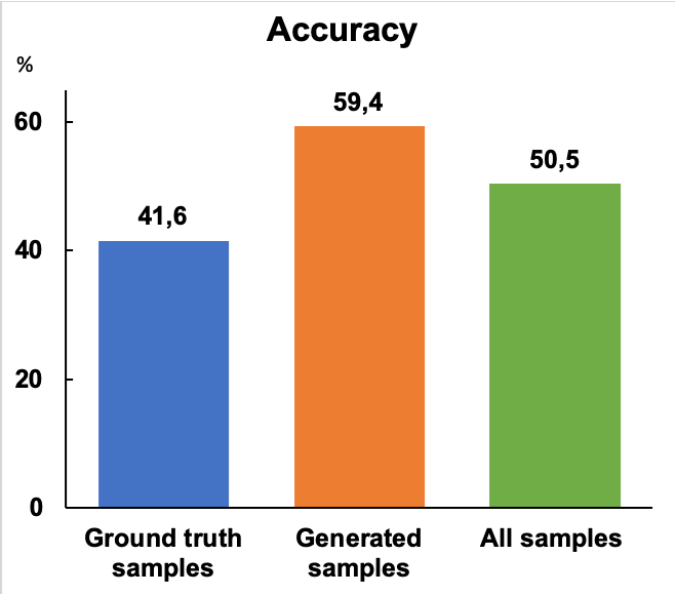


Figure 3.15: The accuracy of all 30 samples, the accuracy of 15 ground truth ones, and the accuracy of 15 generated ones.

By doing this, we would like to see if the participants can classify the robot face. The accuracy of the 30 samples is summarized here, see Figure 3.15. The accuracy is 50.5%, 41.6%, and 59.4%, respectively. The whole samples' accuracy as 50.5% means that subjects cannot recognize the samples well. We can know that the participants cannot identify the generated and real speaking face action during experiments when the generated samples and real samples have similar accuracy. The generated accuracy is 17.8% bigger than the ground truth accuracy, which means that the generated samples are easier to recognize than the real one. All in all, the results indicate that the ground truth face action is better than the generated but the gap is not so big to some degree.

We also explored the accuracy results based on the subjects' personalities (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness), as shown in Figure 3.16. Firstly, no matter what samples (i.E., the real or the generated or the whole samples), the participants with the high score and the ones with low scores have similar accuracy results for each personality trait. For all examples in each personality, all the scores have similar accuracy, about 50%, which reinforces that the generated face action is as good as the real face action to some degree.

Synchronous? Friendly? Human-like?

In this chapter, we also try to compare the generated robot face to the real one for the synchronicity between the speech and the face action, friendliness, and anthropomorphism. Hence, we collected the results of question 2 (Synchronous?), question 3 (Friendly?), and question 4 (Human-like?) during the online experiments as shown in Figure 3.17. We applied the Mann–Whitney U test. We did not use the ANOVA test because the validation scores were not subject to the normal distribution through Shapiro–Wilk test. In this chapter, *n.s.* means $P > 0.05$, $*$ means $P \leq 0.05$, $**$ means $P \leq 0.01$, $***$ means $P \leq 0.001$, and $****$ means $P \leq 0.0001$. All the generated speaking face action is assessed with smaller mean scores than the ground truth for each question from the average rating scores. The mean score difference is not so important. To some degree, it indicates that the synthesized face actions experience less but close performance of synchronicity between the speech and

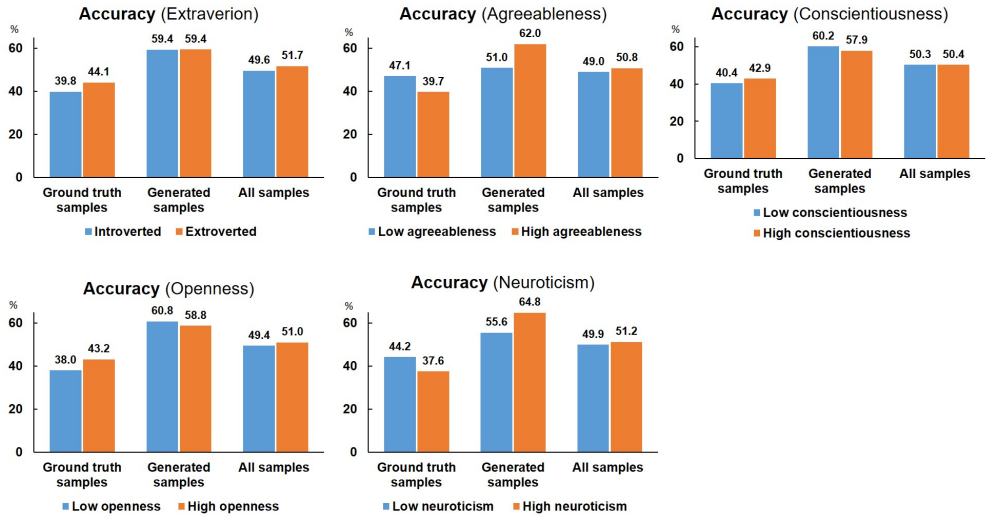


Figure 3.16: The accuracy results based on different five personalities traits

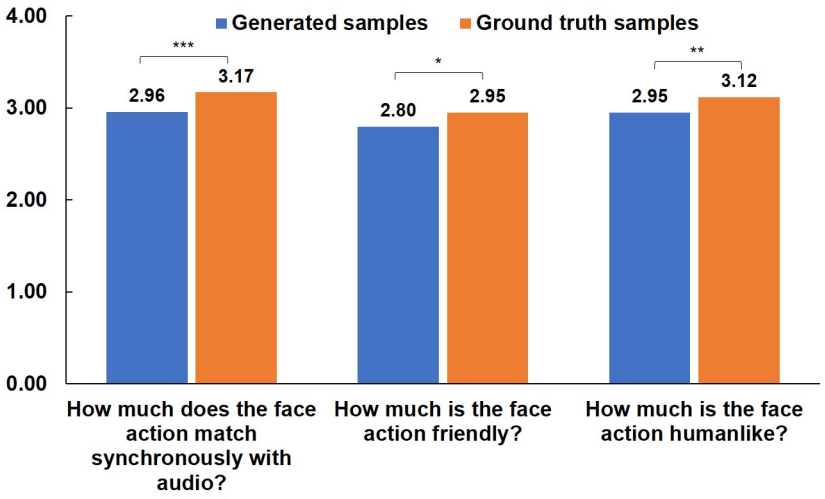


Figure 3.17: The whole assessment results of question 2, 3 and 4.

the face action, friendliness, and anthropomorphism than the real speaking face action. However, based on the statistical significance analysis as shown in Figure 3.17, the ground truth scores are still significantly higher than the generated face actions for synchronicity (question 2: P=0.0002). For friendliness and anthropomorphism, their results have less statistical significance than the result of synchronicity (question 3: P=0.0155, question 4: P=0.0035). It means that there is still some space of improvement for the generated face action. That is because the robot has only five motors to control face action while a human has more than 40 muscles for facial expression. Hence, even the ground truth face action, it can not meet up with human expectation.

For the extraversion personality trait, the assessment results are shown in Figure 3.18. For both generated face actions or ground truth videos, the introverted participants tended to give a higher rating score during face action assessment than the extroverted subjects for all three questions. From the test of significance, the friendliness results for introverted subjects and extroverted subjects have no significant difference (P>0.05). However, other two questions have different effects.

For the agreeableness personality trait, the assessment results are shown in Figure 3.19.

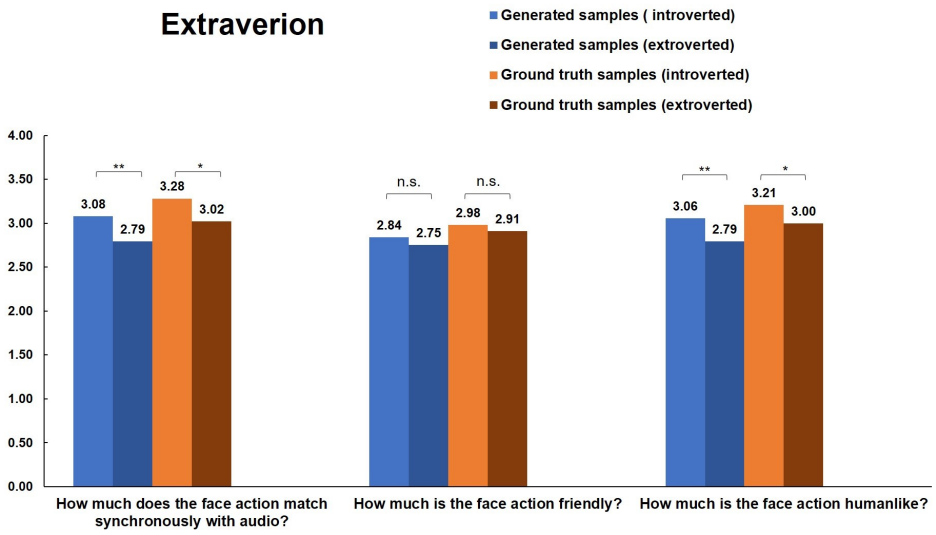


Figure 3.18: The assessment results based on the extraversion personality trait

All the ground truth obtain a little higher mean score than the generated face action with the same parameters. For the second question (friendliness), the subjects with a higher agreeableness score always tend to give a lower assessment score during experiments, and the rating scores of the users with low agreeableness and the users with high agreeableness are not statistically significant (P>0.05). For the friendliness results, the high agreeableness and low agreeableness subjects experience a significant difference in the generated face action assessment. For anthropomorphism, generated face action rating scores have no significant difference (P>0.05) considering the agreeableness personality trait.

For the conscientiousness personality trait, the assessment results are shown in Figure 3.20. The subjects with the higher conscientiousness scores tend to mark a higher score during face action assessment than the subjects with high scores for all three questions. Similarly, all the ground truth face actions obtain a little higher mean score than the generated ones. Different from the assessment of question 3 (friendliness) and question 4 (anthropomorphism), the validation scores of the generated faces have significant differences between low conscientiousness and high conscientiousness ($P \leq 0.05$). Namely, the lower conscientiousness rating score for the generated face actions is significantly higher than the high conscientiousness score of the generated face actions.

For the neuroticism personality trait, the assessment results are shown in Figure 3.21. The users with the lower neuroticism scores tend to mark a higher score during face action assessment than the subjects with high scores for all three questions. Similarly, all the ground truth face actions obtain a higher mean score than the generated ones. However, the statistical significances between lower neuroticism and high neuroticism are different for the three questions. For synchronicity and anthropomorphism, there are significant differences, especially synchronicity.

For the openness personality trait, the assessment results are shown in Figure 3.22. The lower openness scores tend to provide a higher score than those with high scores for all three questions. Similarly, all the ground truth face actions obtain a little higher mean score than the generated ones. However, for all three questions, they have a significant difference (question 2: $P \leq 0.01$, question 3: $P \leq 0.0001$, question 4: $P \leq 0.0001$) during the generated face action validation. It means that personality openness reflects the assessment of the validation scores during experiments.

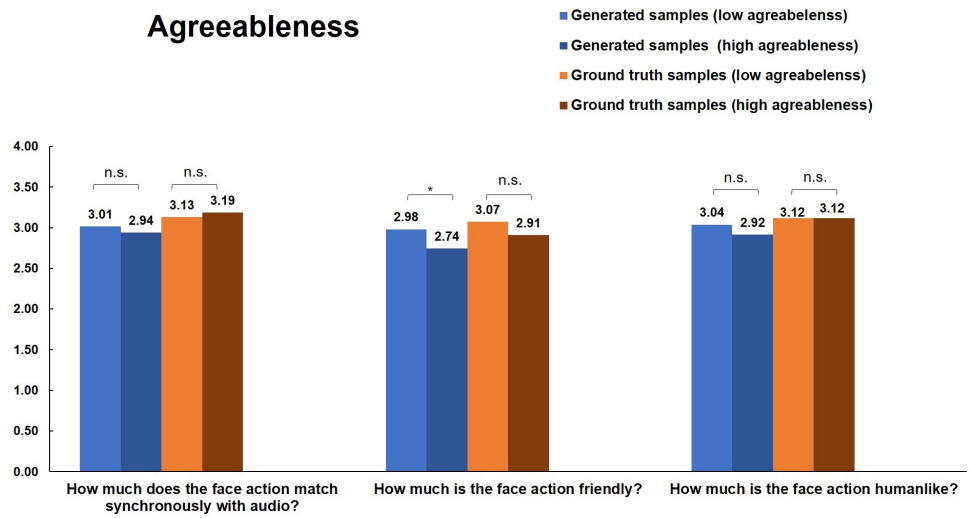


Figure 3.19: The assessment results based on the agreeableness personality trait

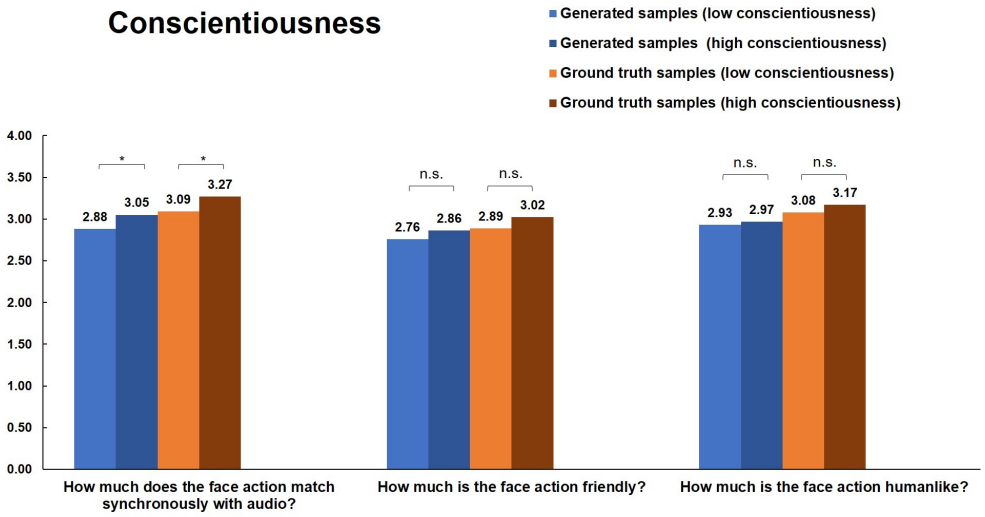


Figure 3.20: The assessment results based on the personality Conscientiousness.

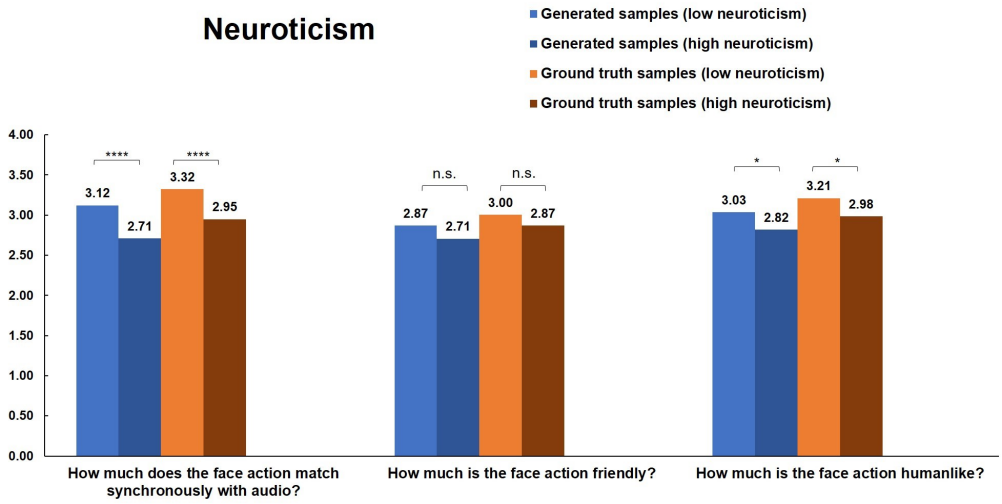


Figure 3.21: The assessment results based on the neuroticism personality Openness.

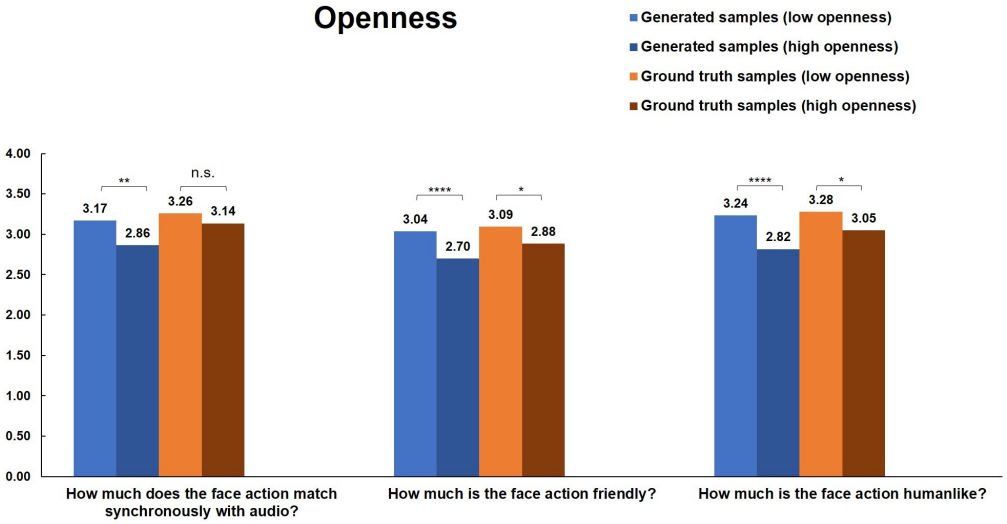


Figure 3.22: The assessment results based on the openness personality trait

All in all, the generated speaking face actions have some space to be improved compared ground truth but is still promising to be used on the human-robot interaction. A robot face with high degrees of freedom for face expression is needed for a higher rating score of the both (the generated and ground truth) to meet up human expectation in the future. The participant personality traits also reflect the validation results during user-joint experiments.

3.7 Summary

The speech-driven robot face action generation also used a GAN-based model to generate the face action from speech audio. However, because speech-to-face mapping (especially the mapping between speech and lip motion) is a more strict alignment task than the speech-to-gesture generation, we did not use one-to-many solution used in the gesture generation. We built an effective temporal GAN architecture, namely *S2FGAN*, with losses of WGAN-GP and L_1 loss for co-speech facial action generation, which is promising to be used for other cross-modal mapping tasks with time series as input and output. The trained *S2FGAN* model can generate realistic and synchronous facial action sequence with speech audio. The facial action retargeting from human face landmarks to robot facial action was completed. Then, we applied the generated a face action based on speech on the face of the Zeno robot. Finally, the generated facial action series were assessed with the qualitative evaluation and the quantitative evaluation. We also conducted user-joined experiments to assess the generated robot face action in comparison to the ground truth. The experimental results show that the generated robot face action and ground truth have no big difference, facilitating a natural human-robot interaction. However, it is still a challenge to get a natural HRI in the wild settings because the Zeno robot has a limited number of motors for facial expression compared with about 40 muscles of humans for facial expression. Hence, more and longer experiments need to be conducted in a natural setup to explore the improved solution in the future. Due to Covid-19 sanitary crisis and the recent lockdown in France, no face to face human-robot interaction experiments could be conducted. In the future, we will do user experiments to explore the long-term human-robot interaction environment with the generated face action presented on Zeno robot. In addition, we will take the emotion label into consideration to explore emotional facial action generation for robot's face.

In this chapter and the previous gesture generation chapter, we only focused on the gesture and face action generation based on speech without considering the user's behavior. A natural step would be the exploration of the generation of the robot's behavior not only based on speech but on the whole human behavior (gesture and facial action). In addition, it would be interesting also to consider the speech text and audio together so as to develop more natural speech-driven robot behaviors. Finally, further research directions could also include the style-controllable speaking robot behavior generation and speech-free robot behavior generation, for example, smile-driven or cry-driven robot behavior (gesture or face action) generation.

3.8 Thesis Contributions

In summary, our contributions in this chapter are as follows:

- We provided a short survey on the generative model for the generative model, state of the art on the face action generation, and the face robot with skin.
- A temporal GAN architecture with L_1 reconstruction loss and GAN loss was proposed to effectively generate a 3D co-speech facial action sequence, which can be possibly used in long-term human-robot interaction. Compared to the speech-to-face mapping

model only with $L1$ loss, our GAN-based model will focus on both global loss and frame-wise loss, which facilitates our model to generate more natural face action instead of the frame-aligned face action by using traditional models. The GAN generator gets the MFCC features of speech audio as input to generate 3D face action. The trained generator can be used for the 3D speech-driven face action generation.

- The facial action retargeting task was performed from human 3D face action to the robot face actuators. The mapping method was used to generate the control signal for the Zeno robot and the Zeno performance certifies the effectiveness of our mapping approach.
- The generated robot facial actions with the related speech were applied to the Zeno robot for human-robot interaction. Furthermore, we asked participants to join the user experiments where Zeno expressed synchronous face actions given the speech to estimate the generated face actions. The results show that our generative GAN model an effective method to generate the speaking face action for a social robot with face skin.

This work is under review [217].

Chapter 4

Multimodal human emotion from thermal face and gait in HRI

4.1 Overview

A human understands the counterpart's emotion and conveys emotion during interpersonal communication. Human emotion perception is an important aspect in social robotics and in human-robot interaction (HRI). A successful human emotion detection during interaction may facilitate a natural human-robot interaction. In this chapter, we propose a vision-based multimodal emotion recognition method based on gait data and facial thermal images designed for social robots. Our method can detect four human emotional states (i.e., neutral, happy, angry, and sad). We gathered data from 15 participants in order to build up the emotion database for training and testing our classification models. We implemented and tested several approaches such as Convolutional Neural Network (CNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), and Random Forest (RF). These were trained and tested in order to compare the emotion recognition ability and to find the best approach. We designed a hybrid model with both the gait and the thermal data and the accuracy of our system shows an improvement of 10% over the other models based on our emotion database. This is a promising approach to be explored in a real-time human-robot interaction scenario.

4.2 Introduction

Emotion understanding plays a very important role in human-human interaction and emotions are strongly related to the social context. Nowadays, robots attempt more and more to understand human internal and emotional states and try to establish more natural social interactions. With the development of more sophisticated sensors and the increase of computational power, more researchers focused on the vision-based methods for emotion recognition. A camera cannot detect a face with a high resolution at high distance but can acquire the gait information during walking. However, at close range, a camera cannot get gait information but can extract clear face images during stand-up position while interacting with the robot. The fusion of the gait data and face images can be used in many situations, including human identification and human emotion recognition. From the perspective of human identification through fusion of gait and face data, Kale et al. [218] used the outdoor data of 30 participants to complete a decision level fusion of gait and face data. However, the gait data was based on the human contours, which were not reliable. Hossain et al. [219] proposed a new multimodal Bayesian method for human identification with features from gait and face data. They applied the PCA-LDA (Principal Component Analysis-Linear Discriminant Analysis) processing in order to get better accuracy. Zhou et al. [220] fused the side face and gait information for human identification. PCA was used to get the features from side face images and gait energy images. In [221], instead of static fusion methods, a context-aware multimodal

fusion method was applied for human identification with gait and face data in real-time contexts. The authors considered two context factors, which reflected the relationship between gait and face during the fusion process, namely the view angle and the distance between the participant and the camera. The context-aware fusion method showed better results with respect to the static fusion rule methods. However, the gait information in these three previous described studies was extracted from 2D images, which provide fewer features than 3D images. From the perspective of emotion recognition with gait or face data, Castellano et al. [222] used the camera to get the data of human body movement and gesture by using the nearest neighbor method with Dynamic Time Warping (DTW) distance to recognize affective states. Mao et al. [112] used the RGB-D data of human face and SVM classifier was employed to recognize the emotions. The authors in [223] presented a method of detecting human emotions with gait data from Kinect. In [224], a real-time 3D facial emotion detection system was presented. The system extracted the dynamic face features and used neural networks and SVR (Support Vector Regressors) to analyze the 16 AU (Action Units) and to complete the facial emotion recognition. However, emotion recognition with RGB-D images or with speech is not always accurate, because humans can lie and hide emotional states. Humans can easily express a facial emotion and feel another emotion. Hence, identifying the real emotional state is crucial.

To the best of our knowledge, no works fuse the gait and the thermal facial features together in order to understand human emotion in a social robotics context. A very common scenario is when the human walks towards the social robot and stops in front of it before starting the interaction. Humans can readily understand the emotion expressed in body action and gait, which are reliable cues for human emotion recognition [225]. Moreover, the thermal facial expression is an important clue for determining the human affective state and thermal data is rarely affected by the change of light exposure condition [226]. In our research, we developed a fusion-method for emotion detection from gait and thermal facial data.

The emotions include the neutral state, happiness, anger, and sadness. The reason why we selected the four emotions is that they are basic emotions. During experiments, we used the Kinect system to extract emotional gait information (i.e., joint angles and joint angular velocities). From the thermal images, thermal features from the ROI are extracted. Additionally, in order to find the most appropriate classifier for the multimodal data, several models including Convolutional Neural Network (CNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), and Random Forest (RF) were trained and tested with gait features and Random Forests were trained and tested with multimodal features gait and thermal facial features. Finally, we compared and interpreted the recognition accuracy of these emotion recognition models and proposed a hybrid model with gait and thermal face data for emotion recognition. Furthermore, we conducted 300 experiments that allowed us to collect RGB-D and thermal data and we evaluated our system. The whole process of multimodal emotion detection is as shown in Figure 4.1.

4.3 State of the Art

In this chapter, our object is to explore human emotion understanding with the fusion of the thermal facial information and the gait features. Hence, in this part, we will review the research on gait-based emotion perception and thermal-based affective computing.

4.3.1 Gait emotion classifier

Gait is related to human motion in daily life. It does not only facilitate human mobility but it also conveys the emotion of the walker. Compared to the biometric interface for emotion

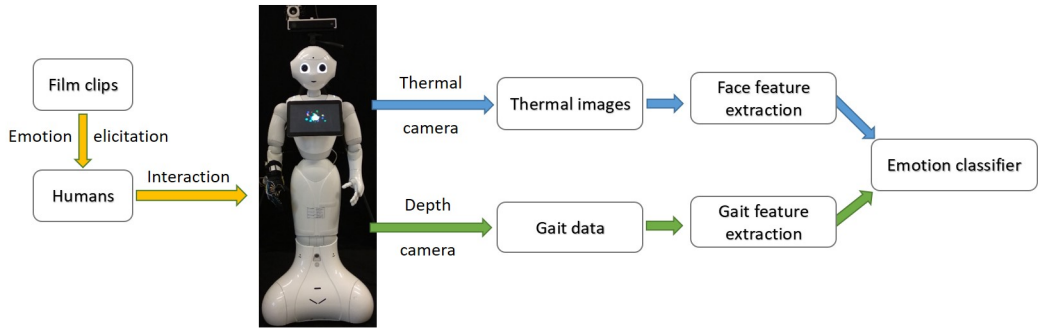


Figure 4.1: Process of emotion recognition. After emotion elicitation, the subject interacts with the robot. During human-robot interaction, the thermal camera and the Kinect sensor record the thermal face images and the 3D gait data, respectively to build up a database. Then, the features are extracted for the multimodal emotion recognition.

recognition based on speech, facial expression, And physiological state, the gait-based affective perception provides a new solution for remotely observable emotion detection. Past works on emotion recognition models have explored the machine-learning methods, for example, the model based on PCA (Principal Component Analysis), the model based on support vector machine (SVM), the classifier with CNN model, and the model with Graph Neural Networks (GNN). Karg et al. [227] applied a PCA-based model for human affective state classification. Authors in [228] used an SVM model with the geometric features, motion features, and Fourier features of human gait to detect four human emotions (i.e., neutral, angry, happy, and sad). The paper [229] built-up a gait emotion recognition classifier based on GNN to detect human emotions. The emotion recognition model was trained on the database recorded from the experiments, and the database augmented through a generative network. The generative model based on autoencoder and GNN can generate emotional skeletal gait samples, which humans annotated for data augmentation. In order to process the spatial and temporal gait features well, the GNN model used the famous architecture of ST-GCN (Spatial Temporal Graph Convolutional Network) [230]. The gait-based emotion classifier network can detect four emotions: happy, sad, angry, or neutral. The paper [231] applied an end-to-end emotion recognition model based on Convolution Neural Network to detect pedestrian emotion. Then, the extracted emotion was used to guide the social robot for proxemic interaction. Video streams with human gait were recorded, and the skeletal poses were extracted frame by frame. The obtained poses were transferred to the image embedding for following emotion recognition.

4.3.2 Thermal emotion classifier

The human face temperature is not only a critical health indicator but also has a strong relationship with the human affective state. Infrared thermography with a thermal camera is a technology that enables us to sense the radiation of energy emitted from the human body. There are many advantages of using infrared thermography: during facial emotion detection, the thermal image is less influenced by illumination and occlusion variation [232] and the thermal camera is contact-free and less invasive than other sensors used to detect the physiological state of human users. The literature review shows a strong relationship between the human facial thermal data and human emotional states (e.g., for happy emotion: [233], [234], [235], [236], for sad emotion: [233], [235], and for angry emotion: [235], [236]). With a thermal camera, the thermal images can be recorded for emotion recognition. Many

research works have been done in the past few years on the human emotion perception with a thermal camera [237]. The authors in ([238]) used a thermal camera to capture the infrared thermography of the face for emotion detection. The paper used a top-down hierarchical system to classify human emotions. The regions of interest, including forehead, cheeks, nose, and maxillary, were detected from the thermal face images through image processing. In order to obtain adaptive emotion recognition for each subject, the self-calibrated system was proposed based on a fuzzy logic model. The model can detect emotions including joy, anger, disgust, fear, and sadness. The paper [239] proposed a fusion model of the RGB face features and thermal face features for facial emotion recognition. The authors explored Bayesian Network (BN) and SVM-based models with decision-level fusion and feature-level fusion. It was found that the fusion method obtained a better accuracy of facial expression recognition than unimodal emotion recognition only with visual features.

4.4 Methodology

Our system includes 5 parts including the emotion elicitation method, the creation of the database, the thermal facial feature extraction, the gait feature extraction, and the emotion recognition algorithm.

4.4.1 Emotion elicitation

In our research, emotion elicitation is used for emotional gait and thermal face data extraction. In lab environments, researchers often use the image-based method, film-based method, music-driven method for emotion elicitation. The authors in [240] showed that static images and film clips are mostly used as stimuli to elicit emotions in laboratory situations. Moreover, in [241] it was shown that the film-clip-based method was the most effective one to elicit emotions. In our work, we also adopted the film-clip-based method to elicit emotions. Before face and gait data extraction, the participants watch some videos extracted from the open film-based affection elicitation database, namely FilmStim [242] for emotion elicitation. This database includes French and English videos. In our experiments, we selected 6 clips (3 in English, 3 in French), 13 clips for happy emotion (3 in English, 10 in French), 16 clips for angry emotion (9 in English, 7 in French), and 17 clips for sad emotion (7 in English, 10 in French). Before starting the interaction with the robot, participants chose the film clips of each emotion randomly for emotion elicitation.

4.4.2 Multimodal database building

Nowadays, researchers have built many databases of thermal face [243] and gait [223] [244], which can be used for human emotion recognition. However, to the best of our knowledge, there is no available database for emotion recognition combining both gait and thermal images. In order to detect human emotion from gait and thermal face features, we built up our own multimodal database with gait and thermal data through user experiments in the lab environment setting.

Experimental design

During the experiment, we used the depth camera of Microsoft Kinect and the thermal camera Optris PI (Optical resolution: 640x480 pixels; temperature range: -20 degree Celsius to 900 degree Celsius) to obtain the gait data and the facial thermal image, respectively as shown in the Figure 4.2. The Kinect SDK can be used for 3D skeleton data extraction and the thermal

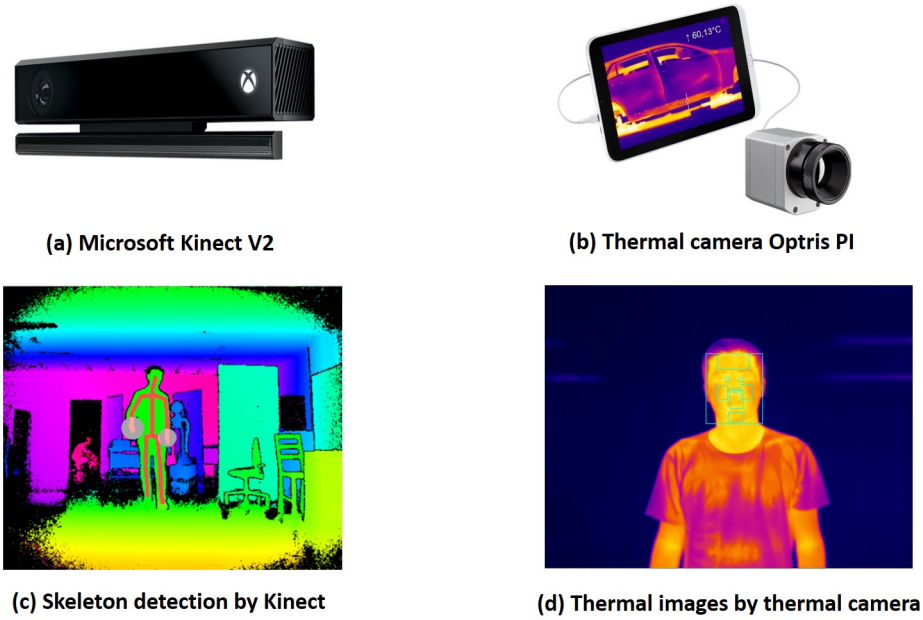


Figure 4.2: Sensors and data extraction. (a) Microsoft Kinect V2 sensor (b) Thermal cameras Optris PI (c) Skeleton detion with Kinect SDk

camera can use the interface of Robot Operating System (ROS) to record the thermal image with face.

In addition, we fit Kinect and thermal camera together through the 3D printed frame. Because the Kinect SDK (Software Development Kit) with skeleton tracking function cannot work on Ubuntu and the thermal camera works well on Ubuntu system with ROS. The Microsoft Kinect sensor works on one computer with the Win10 system and the thermal camera works on another computer with the Ubuntu14 System. Two computers are on the same local network. The gait data from Win10 sends the data to Ubuntu14 so that the Machine learning models could be trained and tested on the same system. The whole experimental setup is depicted in Figure 4.3. Because the thermal camera extracted the thermal images of the upper body and RGB-D camera needed to detect the whole body, the thermal camera and the RGB-D camera had different elevation or depression angle. Firstly, the thermal camera get the thermal images when the participants stand at the focus area of the thermal camera instead of walking to the Kinect. It was easier to set the parameters for the thermal camera than for the Kinect. In our work, we set the focus of the thermal camera at 1.8 meters and its elevation angle at 9.5 degrees upward with the horizontal. We also set the height of the RGB-D camera at 1.27 meters. If the depression angle of the RGB-D camera is 0 degrees, the whole skeleton body data can be obtained only from 2.2 meters to 4.5 meters in the Kinect skeleton axis. In this situation, we cannot obtain enough gait data in each experiment for emotion detection. Hence, we rotated the Kinect angle β downward with the horizontal in order to get enough gait data. When β is 7 degrees, the Kinect working area for whole body skeleton is from 1.7 meters to 4.5 meters and it is suitable for our experiments. Therefore, the angle contained by 2 optical axes of the two cameras is 16.5 degrees. Moreover, the maximum height of the participant who can use the system is 1.95 meters. During the experiment, participants walked from 6m to 1m instead of starting walking from 4.5 meters to 1.7 meters in order to get natural gait information during the experiments. During the experiment, Pepper robot guided verbally the participant to start an emotional interaction. The real sensors and the robot scenario are also shown in Figure 4.3.

All the experiments in this research took place in our laboratory. During the experiments,

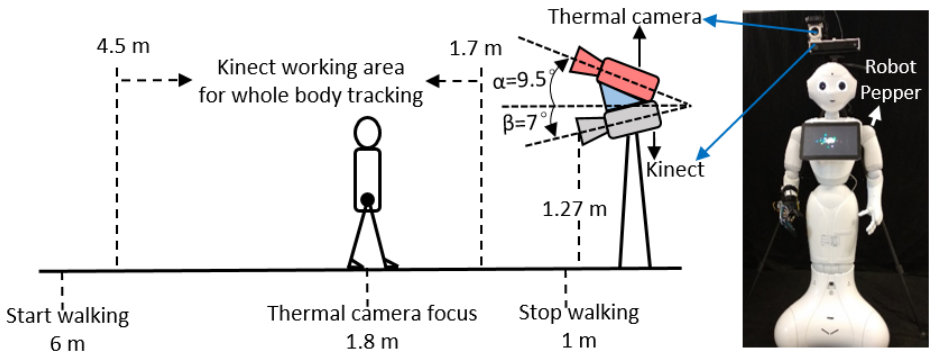


Figure 4.3: Experiment setting

Table 4.1: Emotion sequencing

	1	2	3	4
Order 1	Neutral	Happy	Angry	Sad
Order 2	Happy	Sad	Neutral	Angry
Order 3	Sad	Angry	Happy	Neutral
Order 4	Angry	Neutral	Sad	Happy
Order 5	Neutral	Happy	Angry	Sad

the air conditioning was turned on with the temperature set at 24 degrees Celsius all the time (the face temperature is sensitive to the environmental temperature). All experiments were conducted over a period of 1 month. 15 participants (7 females and 8 males) took part in the experiments for building up the offline database. All of them were students with different major backgrounds and from different universities. Their age ranged from 20 to 32. Each participant joined 20 effective experiments. In the offline experiments, in order to reduce the influence of the emotion order, we randomized the emotion orders based on the size 4x4 Balanced Latin Square (BLS) ([245]). Balanced Latin Square is widely used in experimental designs involving many repeated observations, which can effectively random the experimental orders. In our research, we built the 4x5 Balanced Latin Square as shown in Table 4.1. In the table, there are 5 emotion orders and there are 4 emotions. We added the fifth row as a duplication of the first row in order to get more data from each participant.

During each experiment, the participant firstly watches some film clips for emotion elicitation, and then joins the later experiment for data extraction. In each of the experiment for data extraction, there are two stages including the walking stage and the standing stage. Namely, the participant walks towards the robot, then stops, and stands before the robot to interact for 10 seconds. In the first stage, gait data is extracted from the RGB-D camera while walking towards the robot. In the second stage, thermal images are obtained from the thermal camera during the interaction between the human and the robot. Before the experiment, the participants are also asked to fill in some pre-experiment questionnaires, such as EPQ (Eysenck Personality Questionnaire) for assessing the participants' personality traits and PANAS (Positive and Negative Affect Schedule) for obtaining the participants' positive or negative affective states. The scenario was as follows:

(1) Pepper robot says: "Please position yourself at the starting area";

(2) The participant stands in the area with the sign "Starting line";

(3) Pepper robot says: "Hello! Please walk to me";

(4) The participant walks towards the robot with the emotion elicited by watching the film clips;

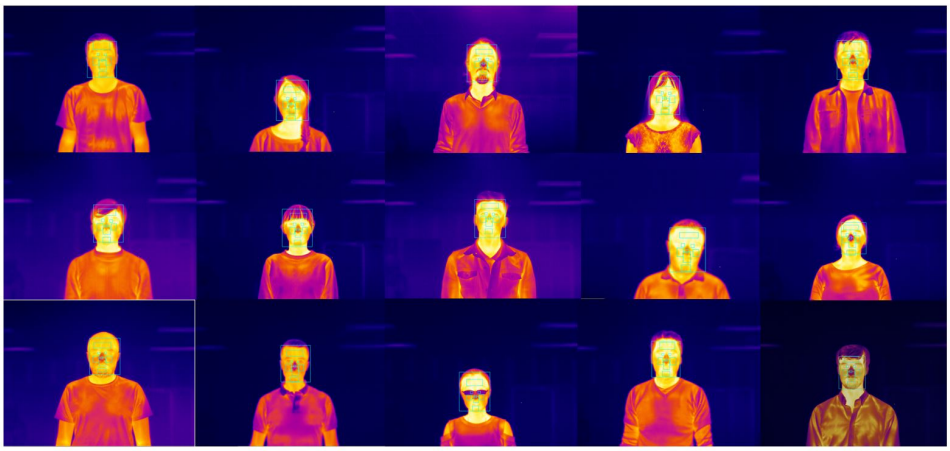


Figure 4.4: Examples of thermal images from the database

(5) The participant stops before the robot and stands there for 10 seconds to interact with the robot.

Data collection

15 participants took part in our experiments. Each of them completed 20 experiments (i.e., 5 neutral, 5 happy, 5 sad, 5 angry) with the random order of emotion experiments. In total, we collected 300 experiments with 75 entries for each emotion. During the experiments, the participants were asked not to move their head much. Some data still contains strong movement of the participant's head. The 5 seconds thermal videos are hand-selected, in which the head of participant does not move or move just a little (the shake of the head leads to inaccurate face ROI detection). Namely, we looked at all the thermal videos one by one and recorded the start time and duration during which participants do not move their heads much. Some examples of thermal facial images included in the database are shown in Figure 4.4.

In the database, we also included the gait data, which corresponds to the positions of the 25 joints that are recorded at 30 fps from the Kinect RGB-D camera. The data from the whole body skeleton was recorded from 0.5 meters to 4.5 meters. The experimental period of the gait recording was about 2-3 seconds. The trajectories of the skeleton for the 4 emotions are shown in Figure 4.5. The purple, green, red, and blue skeletons correspond to the neutral, happy, angry, and sad emotions, respectively. The black trajectories in each one are the trajectories of the head and the base of the spine.

4.4.3 Feature extraction

Gait feature extraction

Human beings can easily understand emotion in gait in daily life. Their understanding is based on the detection of simple features, such as the overall speed of people. As there are so many potential kinematic parameters in human gait, the systematic analysis of gait features is very complicated [225]. In our work, RGB-D images were used to extract the human gait data. We extracted 25 joint positions (x, y, z) of the human skeleton at 30 Hz as shown in Figure 4.6.

Our work used gait data of lower limbs for emotion detection as the lower limbs have more repeatable movements than the upper body as discussed in [246]. Considering future online emotion recognition during human-robot interaction, depth camera on the robot for

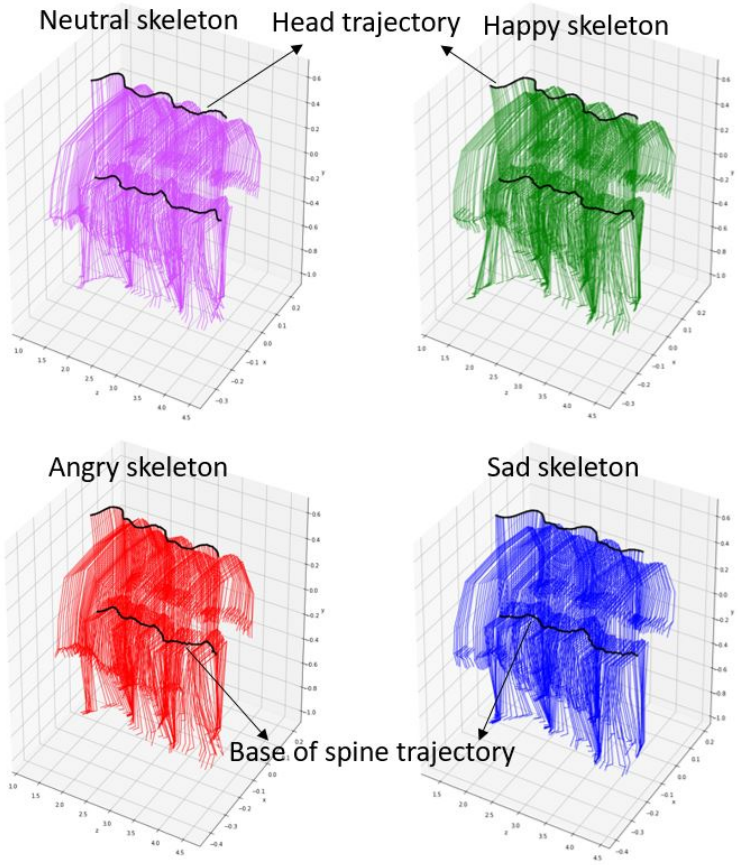
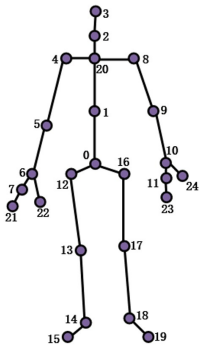


Figure 4.5: The trajectories of the skeleton



(a) 25 skeletal joints

Kinect V2 25 Skeletal Joint Index Infroamtion		
SpineBase = 0	SpineMid = 1	Neck = 2
Head = 3	ShoulderLeft = 4	ElbowLeft = 5
WristLeft = 6	HandLeft = 7	ShoulderRight = 8
ElbowRight = 9	WristRight = 10	HandRight = 11
HipLeft = 12	KneeLeft = 13	AnkleLeft = 14
FootLeft = 15	HipRight = 16	KneeRight = 17
AnkleRight = 18	FootRight = 19	SpineShoulder = 20
HandTipLeft = 21	ThumbLeft = 22	HandTipRight = 23
ThumbRight = 24		

(b) Index infroamtion of 25 joints

Figure 4.6: Human skeleton information extracted from Kinect V2. (a) 25 Skeletal joints (b) Cross reference between index and joints

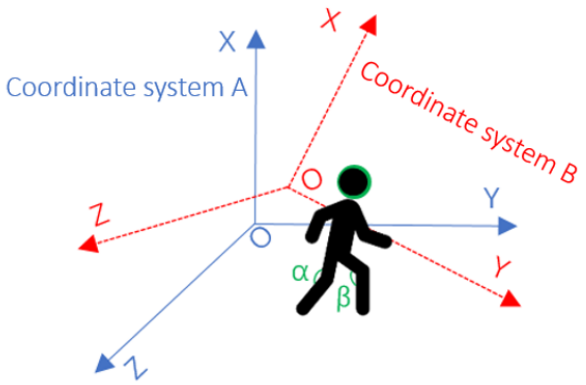


Figure 4.7: Angles in variable coordinate system. The red one is the coordinate system A and the blue one is coordinate system B. The movement of coordinate system on the camera cannot reflect the angular value but it reflects the position values.

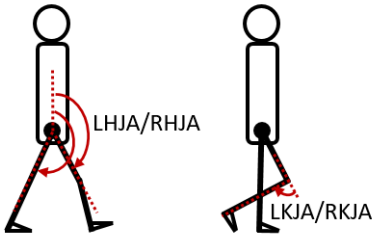


Figure 4.8: Definition of the joint angles

gait data extraction will move all the time. If we use joint position as features, the movement of camera should also be considered, which is very complicated. If the angle is applied as features, we do not need to care about the movement of the camera on the robot. Therefore, here, we selected the joint angle and the joint angular velocity as the features to characterize the gait. As shown in Figure 4.7, even though the coordinate system moves and rotates and the positions (x, y, z) are different in the coordinate system A and coordinate system B, the angles maintain the same value. Hence, we do not need to take the movement of camera situated on the robot into consideration when we use the joint angles as features instead of position.

In our work, 8 gait features were identified as follows: the left knee joint angle (LKJA) and its velocity (LKJAV), the right knee joint angle (RKJA) and its velocity (RKJAV), the left hip joint angle (LHJA) and its velocity (LHJAV), the right hip joint angle (RHJA) and its velocity (RHJAV). The definition of the joint angles is as shown in Figure 4.8.

LHJA, LKJA, RHJA, and RKJA are calculated with the joint position (x_i,y_i,z_i), "i" is the joint index (from 0 to 24) (see Fig. 4.6). LHJA is extracted from the positions of joints 1, 0, 12, and 13. LKJA is extracted from the positions of joints 12, 13, and 14. RHJA is obtained from the positions of joints 1, 0, 16, and 17. RKJA is got from the positions of joints 16, 17, and 18. LHJAV, LKJAV, RHJAV, and RKJAV are calculated through a subtraction operation of the current angle and the previous angle and a following division operation on the sampling time gap, namely 1/30 second. The procedure of angle calculation is as follows. Firstly, the 3D vector $V(i,j)$ is obtained based on the joint i position and joint j position in (4.1). Then, the length of the vector and the inner product of 2 vectors are calculated as shown in equations (4.2) and (4.3), respectively. Equation (4.4) is used for the calculation of the angle between 2 vectors. LHJA, LKJA, RHJA, and RKJA are calculated as shown in equations

(4.5), (4.6), (4.7), and (4.8), respectively. For LKJA and RKJA, we directly use the angle between 2 vectors as the joint angle. For LHJA in (4.5) and RHJA in (4.7), based on whether the thigh swings before or after the torso, there are two different angle calculation methods, respectively.

$$V(i,j) = ((x_i - x_j),(y_i - y_j),(z_i - z_j)) \tag{4.1}$$

$$D(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{4.2}$$

$$VM(i,j,m,n) = V(i,j) \cdot V(m,n) \tag{4.3}$$

$$A(i,j,m,n) = arccos(\frac{VM(i,j,m,n)}{D(i,j) * D(m,n)}) \tag{4.4}$$

$$LHJA = \begin{cases} 180 - A(1,0,12,13), & \text{if } z_{13} < z_0; \\ 180 + A(1,0,12,13), & \text{otherwise.} \end{cases} \tag{4.5}$$

$$LKJA = A(12,13,13,14) \tag{4.6}$$

$$RHJA = \begin{cases} 180 - A(1,0,16,17), & \text{if } z_{16} < z_0; \\ 180 + A(1,0,16,17), & \text{otherwise.} \end{cases} \tag{4.7}$$

$$RKJA = A(16,17,17,18) \tag{4.8}$$

Furthermore, in order to find the most appropriate model for gait and thermal face features, we make use of different emotion classification methods, including CNN, HMM, SVM, RF, and a hybrid model with RF, which needs distinctive gait features. For gait features, CNN and HMM directly use the angles and the angular velocities and the others use Power Spectral Density (PSD). PSD shows the energy of the specific frequency or the frequency range, which is a very useful feature for gait data. Welch method [247] can be used to obtain the PSD features (this is widely used in the literature [248] [249]). In our study, we use it for PSD extraction. The steps are as follows:

1. Divide the sequence of the joint angular velocity or the joint angle - g[0], g[1], g[2],..., g[N-1] into M segments.

   segment 1: g[0], g[1],..., g[H-1]

   segment 2: g[S], g[S+1],..., g(S+H-1)

   ...

   segment M: g[N-H], g[N-H+1],..., g[H-1]

   where,

   N: Size of the gait sequence

   M: Number of segments

   H: Length of every segment

   S: Number of values to shift between neighbor segments.

2. Calculate the window-based discrete Fourier transform (DFT) of each gait data segment and then get the related periodogram values $P_m(f)$from the results of DFT.

$$w(h) = sin^2(\frac{\pi h}{L-1}) \tag{4.9}$$

$$W = \sum_h w^2(h) \tag{4.10}$$

$$G_m(f) = \sum_l g[h] * w[h] * e^{-2\pi jfh} \tag{4.11}$$

$$P_m(f) = \frac{|G_m(f)|^2}{W} \tag{4.12}$$

$$f = \frac{q}{h}, 1 - \frac{h}{2} \leq q \leq \frac{h}{2} \tag{4.13}$$

where,

f: frequency of DFT

m: Digit of each segment $(1 \leq m \leq M)$

$w(h)$: window function with Hanning function [250]

$h : (m-1)S \leq h \leq H + (m-1)S - 1$

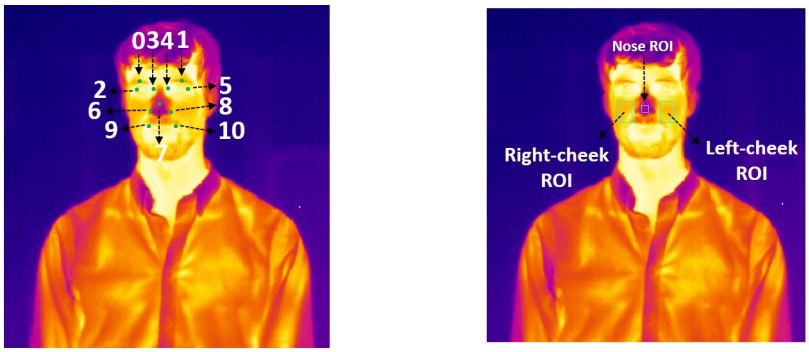
3. Obtain the PSD

$$PSD = \frac{1}{M} \sum_{m=1}^{M} P_m(f) \tag{4.14}$$

Thermal face feature extraction

After obtaining the thermal facial images from the thermal camera, we will do feature extraction. Some recent works show that the nasal and perinasal areas are the most sensitive areas to the temperature change in the different emotional states and the temperature increases with the positive emotions and decreases with the negative emotions [251]. In addition, the left-cheek and right-cheek areas are linked to the changes in the emotional state [234] [251] [252]. Based on these findings, we selected three facial areas as regions of interest (ROI) in order to understand human emotions, namely nose, left-cheek, and right-cheek areas.

During each experiment, the average temperature values and the variances of these three ROIs were calculated. Facial key points in the thermal image were detected through the key point detector based on the Dlib face detection interface [213]. The face key point detector was trained with 493 images from the Natural Visible and Infrared facial Expression Database (NVIE) [243] and from our previous experiments. The trained model can detect 11 key points (i.e., the middle point of the eyebrows, the inner and outer corners of the eyes, the corners and the tip of the nose, and the corners of the mouth), as shown in Figure 4.9(a). The nose region is a square region around the tip of the nose (namely key point 7) and the sides are equal to 1/3 of the distance of eyes. Here, the distance of eyes is equal to the distance between the key point 3 and the key point 4. The left cheek and right cheek regions are rectangular regions with the width equal to the length between the corners of the eyes, and the height equal to the half of vertical distance between the corner of the mouth (key point 9 or 10) and the corner of the eyes (key point 2 or 5). More details are described in the paper [253](b). The 3 ROIs of the thermal face are transferred from 11 key points as shown in Fig. 4.9.



(a) 11 face key points (b) 3 face ROIs

Figure 4.9: Thermal face feature extraction. (a) 11 face key points detected through *Dlib* face detection interface. (b) The three face ROIs including the left-cheek ROI, the right-cheek ROI, and nose ROI.

4.4.4 Emotion classification algorithm

Nowadays, there are a lot of machine learning algorithms that can be used for human emotion recognition. In order to find the most appropriate classification approach for the gait and thermal face features, we implemented CNN, HMM, SVM, and RF to train and test our multimodal database. We also developed a new hybrid classification model based on the best unimodal model above with both gait and thermal facial features. CNN and HMM only used the gait data as time series. SVM and RF used all the features including thermal and gait features. The input size of CNN was [8,56] where 8 is the number of time series. HMM, SVM, RF, and our hybrid model used an input of [1,126]. Namely 120 values as features extracted from 8 gait time series (15 values from each time series) and 6 values extracted from ROIs of the thermal face images.

Unimodal emotion classification model

Convolutional Neural Network (CNN) is one of the most popular deep learning methods, due to its powerful modelling capabilities, which is often applied to image and video processing [254]. In our emotion classification model with 1D CNN, there were 8 gait features corresponding to 8 time sequences, namely LKJA, LKJAV, RKJA, RKJAV, LHJA, LHJAV, RHJA, and RHJAV. The sequences build a matrix with 8 rows and 56 columns (56 is the length of every gait time sequence). The working range of Kinect for 3D gait data extraction is from 1.5 meters to 4.5 meters. Hence, the average time duration of the walking part is for about 2 seconds. The frame rate of Kinect for gait data extraction is 30 Hz. There were about 60 frames in each experiment and the shortest were 56 frames. In order to avoid the padding with zero frame, we selected the first 56 frames from each experiment for emotion recognition. The matrix [8,56] corresponds to a pixel value matrix of the gray-scale image in the CNN model for image processing. Then, similarly, CNN can be used for time sequence classification just like the image part. CNN architecture for gait emotion classification is shown in Figure 4.10. In the CNN model, convolution kernel size is [1, n], which means that convolution operation goes on respectively in each row instead of among rows and columns during image convolution. The images have 2D structure, which are highly correlated spatially while gait time series have a strong 1D structure, which is highly correlated temporally [12]. Zero padding in the sequence matrix is used during convolution operation. In addition, the activation function is ReLU (Rectified Linear Unit). Then, we selected the max pooling during pooling operation. The pooling operation only goes on in each row of the input

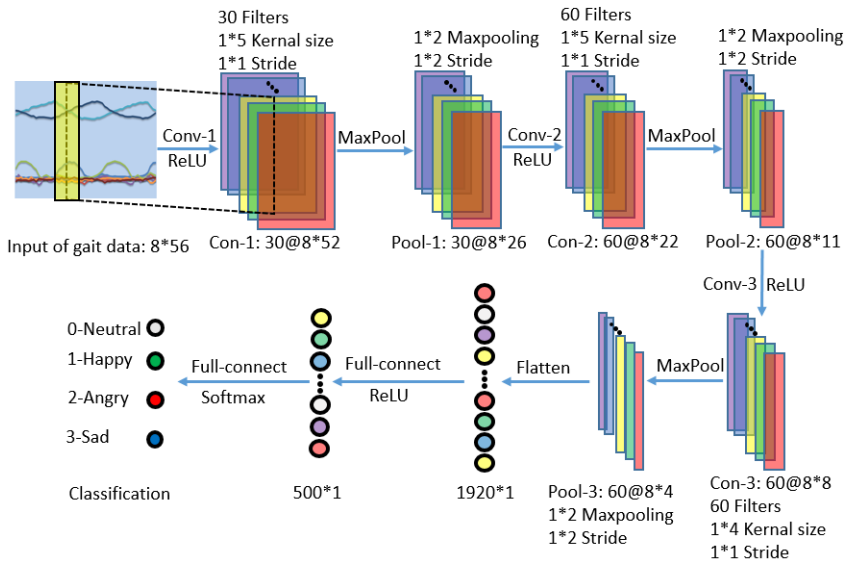


Figure 4.10: 1D CNN model with gait data as time series.

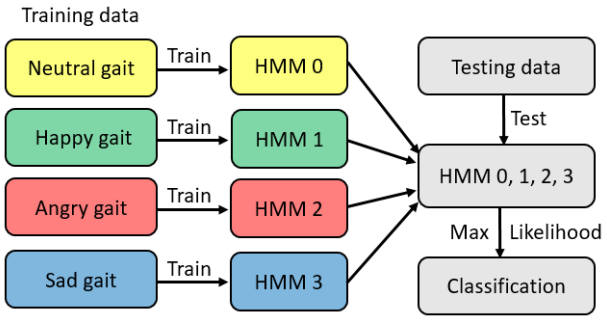


Figure 4.11: HMM model with the gait PSD features.

sequence matrix. Afterwards, the feature matrix is flattened and follows a fully-connected layer with a softmax function.

Brand et al. [255] presented the algorithm with Hidden Markov Model (HMM) to classify the two-handed actions during T'ai Chi Ch'uan and the action classifier shows a good accuracy of action recognition. However, human emotion recognition in action is different with respect to action recognition as the emotional state is more complicated to recognize being highly dependent on the inner state of the human. In our work, we apply HMM to classify the 4 emotion states and the model structure is as shown in Figure 4.11. In the classifier model, four HMM models were trained, respectively with the corresponding emotion training database at first. The recognition process was based on comparing the likelihoods of testing data from the 4 HMM models and choosing the maximum one as the recognition result.

Support Vector Machine (SVM) model and Random forest (RF) dealt with non-linearly separable data effectively as the emotional gait data is high-dimensional and difficult to classify. In this chapter, we applied SVM and RF, respectively with gait data or thermal facial data for emotion detection in order to find the best model for the social robot. In addition, two of them utilized the Power Spectral Density (PSD) as gait features. The structure of the emotion classification models is shown in Figure 4.12. In the SVM model, the kernel is

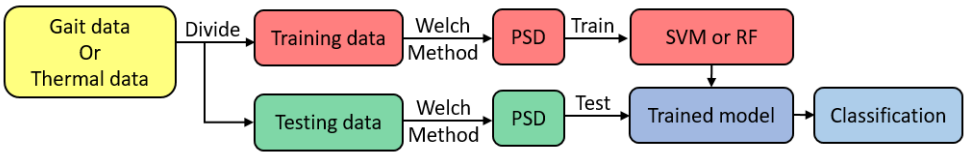


Figure 4.12: SVM model and RF model with gait PSD features.

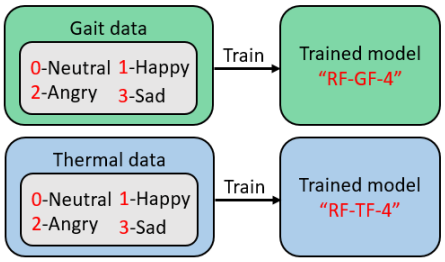


Figure 4.13: Unimodal RF models used in the hybrid model for four emotions.

Radial Basis Function (RBF), which function is defined as shown in Equation 4.15:

$$K(x,x') = exp(-\frac{||x-x'||^2}{2}) \tag{4.15}$$

Multimodal hybrid classification model

After testing the classifiers above only with the gait features from our database, we obtained the accuracy of the models. The CNN, HMM, SVM, and RF have a low accuracy. In order to improve the emotion classification, we fused the gait features and thermal facial features to build up a hybrid model with Random Forest (i.e., RF was chosen because as shown in Table 4.3, it obtained a better performance on the emotion classification with gait features than SVM, HMM, CNN during our experiment). The steps of building up the hybrid model are detailed below.

Firstly, we trained one RF model only with the thermal facial features of the four emotions and another RF model only with the gait features of the four emotions. The related process is shown in Figure 4.13. The model "RF-GF-4" means the RF model with gait features for the four emotions. The model "RF-TF-4" means the RF model with thermal facial features for the four emotions.

Secondly, we trained other six RF models to classify two emotions, respectively. Every model used both the thermal facial features and the gait features of two different emotions. The models are shown in Figure 4.14. For example, the model "RF-TGF-01" is trained with both gait and thermal features of Emotion 0 and Emotion 1 (where, Emotion 0: Neutral, Emotion 1: Happy, Emotion 2: Angry, Emotion 3: Sad).

Lastly, we used the eight models trained above to build up the hybrid emotion classification model as shown in Figure 4.15. At first, the gait features are used as input of the model "RF-GF-4" to get the output "Emotion m" when the thermal one is input the model "RF-TF-4" meanwhile to get the output "Emotion n" (m, n = 0, 1, 2, 3). If Emotion *m* equals Emotion *n*, it means that the two models have the same classification result. In this situation, Emotion *m* or Emotion *n* is the final result. If the two outputs are not equal, the features will be input into the model "RF-TGF-mn" to get the output, which result is thought of as the final human emotion detection result. In this situation where the two outputs are unequal, it

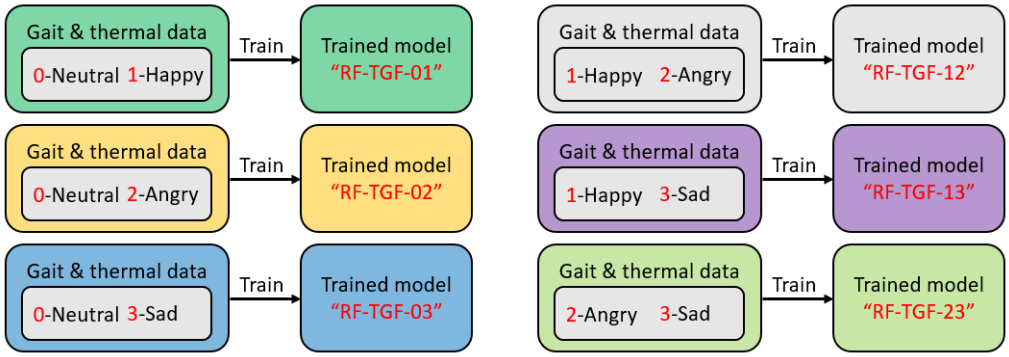


Figure 4.14: RF part with 2 emotions in the hybrid model

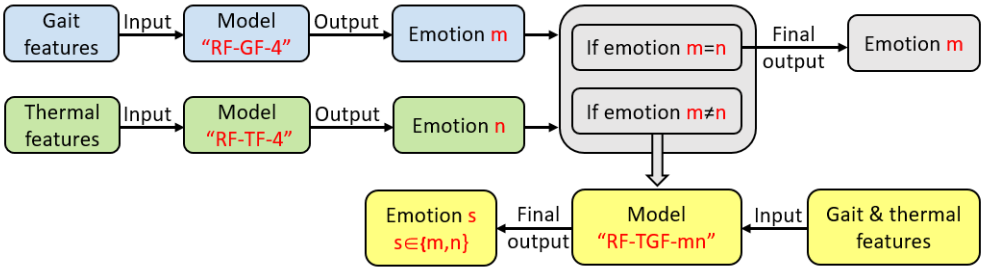


Figure 4.15: The structure of hybrid model

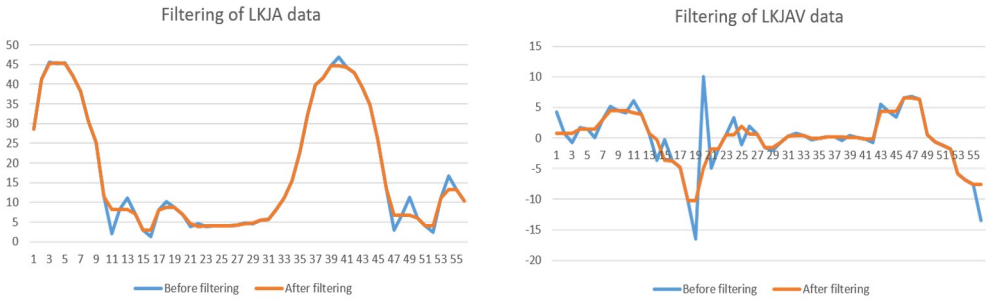
means that the two single RF models classify the input data into two emotions with greater possibility as true emotion. In addition, the classification of two emotions is easier than the one of 4 emotions, and hence, the model can potentially obtain a better accuracy.

4.5 Data analysis

The hybrid emotion detection model used the gait data and the thermal facial images to classify the four emotions. Here, we analyze and process the two data flows in order to build up a better hybrid model for emotion detection. For the gait data, the data processing includes the median filtering and calculation of Power Spectral Density (PSD). Firstly, the median filter as a non-linear tool is used to remove the noise from the eight gait sequences. For example, Figure 4.16(a) and Figure 4.16(b) show the filtering results of LKJA and LKJAV. After the filtering process, the sequence is smoothed, which can effectively reduce the reflection of the noise during the machine learning process.

Secondly, PSD is calculated by using the Welch method and 15 frequencies is uniformly selected from 0 Hz to 50 Hz. The PSD results of one sample for eight time series are shown in Figure 4.17.

Figure 4.18 describes the distribution of the mean in each ROI in the 8 experiments from one participant. It can be seen in the figure that the different emotions have different distributions. The ROI average temperature and variance of 8 experiments from one participant are shown in Figure 4.19. The different four emotions show different feature patterns in the Figure. From Figure 4.18 and Figure 4.19, we can get that the four emotions can accurately be classified through the thermal features above.



(a) Filtering of LKJA (b)Filtering of LKJAV time series

Figure 4.16: Gait data filtering with median filter. (a) filtering of LKJA (b) filtering of LKJAV

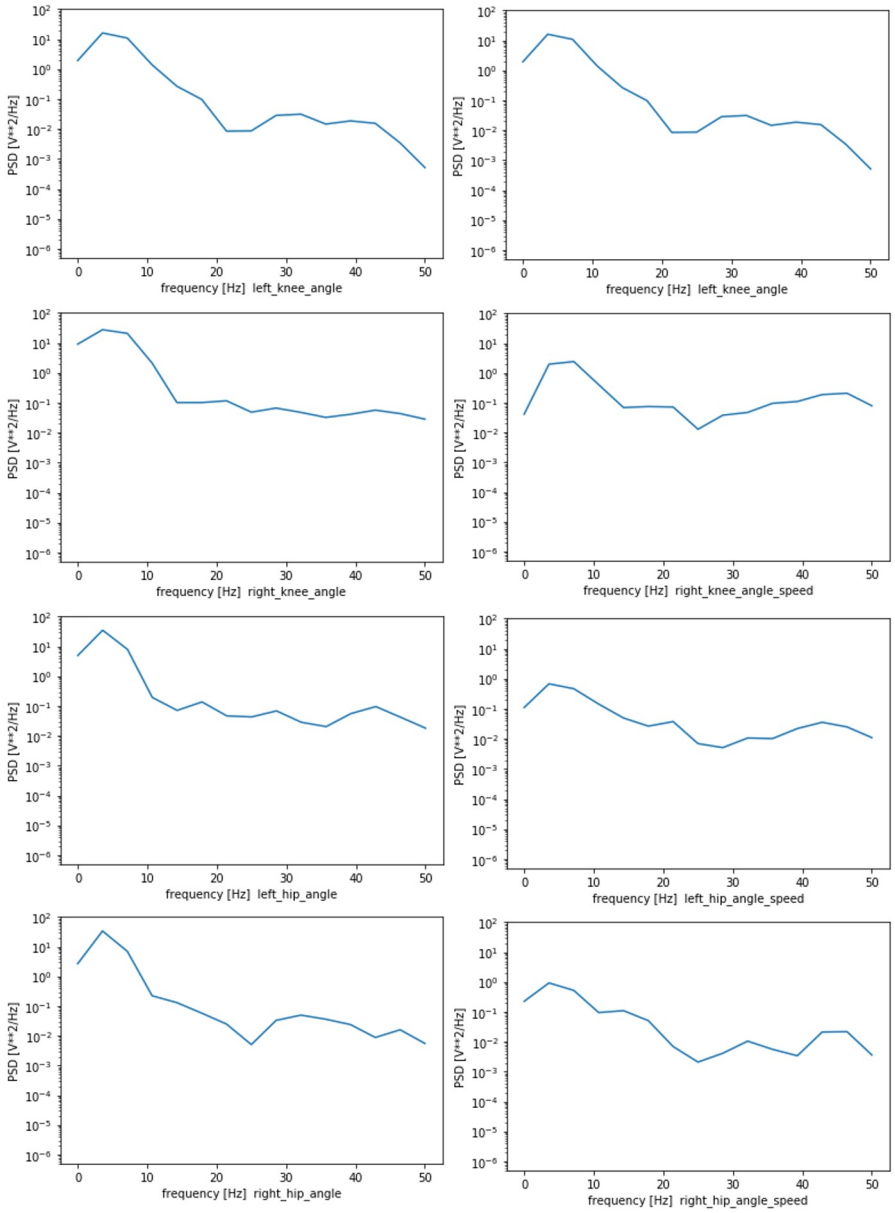


Figure 4.17: PSD of 8 gait sequences for one sample in database.

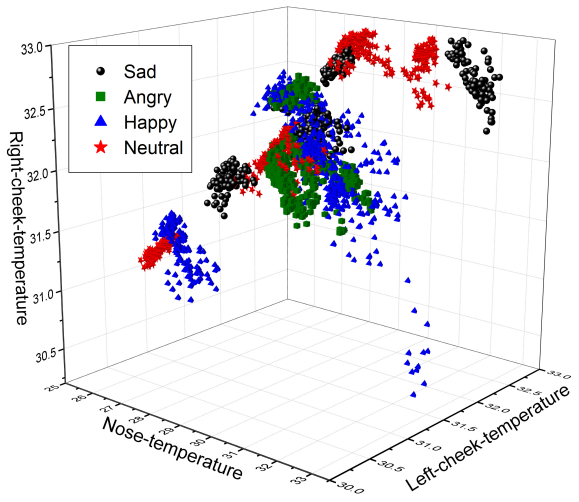


Figure 4.18: Temperature distribution of 3 ROIs with different emotions (black circle: Sad, green cube: angry emotion, blue tetrahedron: happy emotion, red pentagram: neutral state). The four emotions have different distributions.

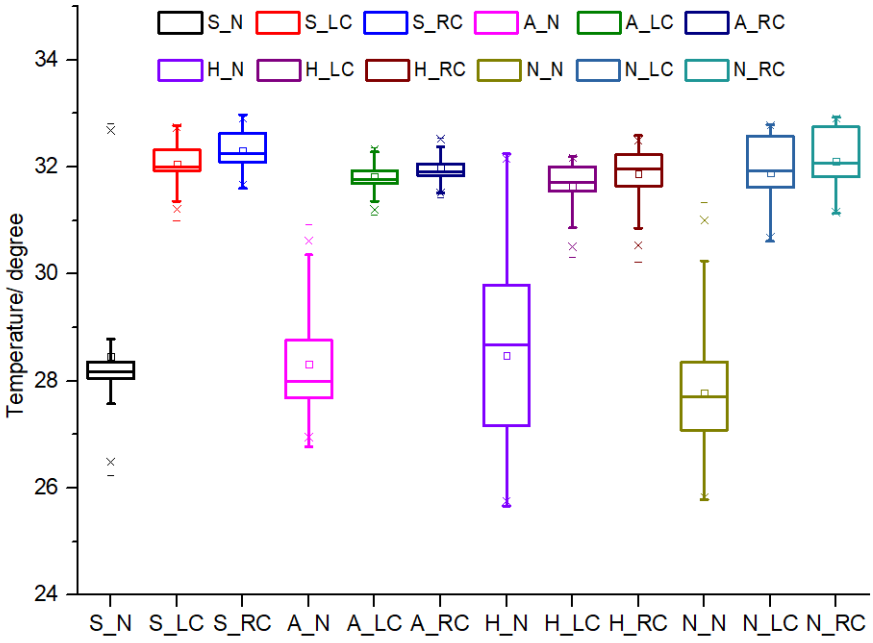


Figure 4.19: Mean temperatures and temperatures variance of 3 ROIs with different emotions for one participant (N-: Neutral, H-: Happy, A-: Angry, S-: Sad, -LC: Left Cheek, -RC: Right Cheek, -N: Nose). The four different emotions show the different feature patterns.

Table 4.2: Accuracy of RF with 2 emotions

Model	Accuracy	Model	Accuracy
RF_01	90%	RF_12	90%
RF_02	80%	RF_13	90%
RF_03	80%	RF_23	70%

4.6 Training and testing results

In our hybrid model, the RF models with both features for two kind of emotions were applied. There were 6 models in total and the testing results of the 6 models are shown in Table 4.2. The model "RF_mn" means Random Forest model that only classifies 2 emotions, namely the emotion m and the emotion n (m, n = 0, 1, 2, 3; 0: Neutral, 1: Happy, 2: Angry, 3: Sad). From the table, it can be observed the high classification scores of 2 emotions with "RF_mn". In the first part of our hybrid model, the RF model with the thermal features and the RF model with the gait features get the classification results, respectively. In the second part, the hybrid model compares the 2 classification results. If they are the same, the same value is kept as the final result. Otherwise, based on the 2 different results, all the features are used as input towards one related model of 6 RF models only for 2 emotions' classification. The testing results of the 6 models are shown in Table 4.2. The model "RF_mn" means Random Forest model that only classifies 2 emotions, namely the emotion m and the emotion n (m, n = 0, 1, 2, 3; 0: Neutral, 1: Happy, 2: Angry, 3: Sad).

The offline-testing results of our hybrid model are summarized in Table 4.3. And training size: validation size: testing size is 12:2:1. At first, by comparing Tables 4.2 and 4.3, we can find that the classification of 2 emotions is better than the one of 4 emotions. So we used the hybrid model to classify the emotions into one (e.g., the two results of the model "RF-GF-4" and the model "RF-TF-4" are the same) or two results (e.g., the two results of "RF-GF-4" and "RF-TF-4" are different). Then, if there are 2 emotions, the hybrid model will adopt the "RF-TGF-mn" in the second part to finish the classification. From Table 4.3, the end-to-end CNN model with the gait data as a deep learning method has only an offline testing accuracy of 55%. The offline accuracy of HMM model with the gait data and without PSD features is only of 65%. SVM model and RF model only with the gait PSD features have the offline testing results of 65% and 70%, respectively. And, SVM and RF models only with thermal features classify the 4 emotions with the accuracy of 55% and 60%, respectively. SVM with both kinds of features only gets 70%. Our hybrid model has the highest offline accuracy, 80%.

We also analyzed the data based on gender. With our hybrid system, we found that the emotions expressed by female participants were better recognized than the ones done by male participants with an accuracy of 80% and 65%, respectively. This is in line with the literature reviews that show small but significant gender differences in emotion expressions, with females showing greater emotional expressivity [256], [257]. We also computed the confusion matrix of the emotion recognition to check the difference of classification results of different emotions. The confusion matrix is shown in Table 4.4. We can see that the happy emotion is the easiest one to classify.

4.7 Summary

In this chapter, firstly, we finished our multimodal experiment settings with various hardware platforms (i.e., thermal camera, depth camera, and the Pepper robot), the multimodal

Table 4.3: Accuracies of different classification models

Model	Data Source	Accuracy
CNN	Gait	55%
HMM	Gait	65%
SVM	Gait	65%
SVM	Thermal	55%
SVM	Gait&thermal	70%
RF	Gait	70%
RF	Thermal	60%
Hybrid model	Gait&thermal	80%

Table 4.4: Confusion matrix of 4 emotions

Confusion Matrix		Predicted class			
		Neutral	Happy	Angry	Sad
Actual class	Neutral	80%	0%	0%	20%
	Happy	0%	100%	0%	0%
	Angry	20%	20%	60%	0%
	Sad	0%	20%	0%	80%

data extraction interface and the emotion elicitation part. User experiments were conducted to build up a multimodal database for emotion classification. In order to explore our new multimodal database, we tested unimodal emotion classifiers such as 1D CNN, HMM, SVM, and RF with gait and thermal facial features of samples in our database. We also proposed a new hybrid model with thermal face features and gait features to classify four human emotional states (i.e., neutral, happy, angry, and sad). The results show that our hybrid model performs better than the other unimodal classifiers (i.e., 1D CNN, HMM, SVM, RF). Even if our results are encouraging, a larger database is needed in order to explore the performance of our model. Furthermore, we plan to conduct a new experiment with the Pepper robot in a natural interaction to test the accuracy of our method and how the trust of the human users is influenced based on the recognition rate.

4.8 Thesis Contributions

In summary, our contributions in this chapter are as follows:

- Based on the HRI experiments, we built up a multimodal emotion recognition database that contained the thermal face images with long-term emotion features and the 3D gait data with short-term emotion features.
- We explored the feature extraction method for 3D gait data and thermal facial images. The PSD-based feature extraction method was used in our research. The face key points were detected from thermal facial images and then were used to obtain the ROIs such as left-cheek ROI, right-cheek ROI and nose ROI. These ROIs' statistical temperature features are used for emotion recognition.
- We tested multiple unimodal emotion classifiers such as 1D CNN, HMM, SVM, and RF on our database.
- We came up with a hybrid model for multimodal emotion recognition, which was also tested on our database and has a lot of potential for an online natural HRI.

This work has been published at [142].

Chapter 5

Interactive robot learning for multimodal emotion recognition

5.1 Overview

Human emotion recognition is critical for a natural human-robot interaction. Most of the methods used for human emotion recognition are offline. In this chapter, we present a long-life multimodal emotion recognition system during human-robot interaction by using an interactive robot learning method with human in the loop. The framework uses two types of features extracted from thermal facial data and 3D gait data. Furthermore, the human verbal feedback is injected into the robot long-life learning process for the multimodal emotion perception. The first results show that our method is a promising solution to be explored in a real-time human-robot interaction scenario.

5.2 Introduction

Based on the participants' experiments, we have collected RGB-D data for gait and facial thermal data in order to build-up our offline multimodal emotion database as discussed in *Chapter 4*. We have conducted the offline testing on our multimodal database. However, the online emotion recognition is still challenging during the real-world human-robot interaction. This chapter explored how to find a solution to overcame this problem with human in the loop with the multimodal emotion recognition model and how to build up a hybrid model with multimodal information for long-term emotion recognition.

In order to find the most appropriate classifier with unimodal data as basic model of the hybrid model for the multimodal data, several models including Convolutional Neural Network (CNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), and Random Forest (RF) were trained and tested. The unimodal classifiers for offline emotion recognition have been detailed and discussed in *Chapter 4*. After comparing the obtained accuracy for each model, we selected the best one as basic model to build up the hybrid model with decision-level fusion with the modified confusion matrix.

Even though many researchers study the multimodal emotion recognition models for robots, the online testing of emotion recognition model is challenging in real-time HRI context. This is mainly due to the related emotion databases that have limited size and because it is very expensive and time-consuming to label and annotate large databases by hand. Interactive Robot Learning (IRL) with human in the loop can overcome this problem. In human-robot interaction, robots can obtain the verbal feedback from humans to label or relabel the data extracted from the interaction ([258]). IRL is very useful in the long-life learning context where there is no large-scale data for emotion recognition. Robots can record the emotion-related features and obtain its label from the interaction with humans. If the online emotion classification gives a mismatched answer, the robot can record the data of the interaction and

relabel it to retrain the model. In this chapter, the interactive robot learning experiments were performed by using the human vocal feedback in the learning loops. Based on the received feedback, the emotion recognition model decided whether the data should be relabeled and the model was retrained at each loop. The hybrid model learned from the interaction and updated the modified confusion matrix for the next online testing experiments. In addition, the online testing was performed again to check the improvement of the hybrid model with the help of the interactive robot learning. Finally, we observed an improvement of the accuracy of the online testing, which shows the effectiveness of the interactive robot learning in multimodal emotion recognition.

5.3 State of the Art

From the perspective of human identification through fusion of gait and face data, Kale et al. ([218]) used the outdoor data of 30 participants to complete a decision level fusion of gait and face data. However, the gait data was based on the human contours, which were not reliable. Hossain et al. ([219]) proposed a new multimodal Bayesian method for human identification with features from gait and face data. They applied the PCA-LDA (Principal Component Analysis-Linear Discriminant Analysis) processing in order to get better accuracy. Zhou et al. ([220]) fused the side face and gait information for human identification. PCA was used to get the features from side face images and gait energy images. In ([221]), instead of static fusion methods, a context-aware multimodal fusion method was applied for human identification with gait and face data in real-time contexts. The authors considered two context factors, which reflected the relationship between gait and face during the fusion process, namely the view angle and the distance between the participant and the camera. The context-aware fusion method showed better results with respect to the static fusion rule methods. However, the gait information in these three previous described studies was extracted from 2D images, which provide fewer features than 3D images.

From the perspective of emotion recognition with gait or face data, Castellano et al. ([222]) used the camera to get the data of human body movement and gestures by using the nearest neighbor method with Dynamic Time Warping (DTW) distance to recognize affective states. Mao et al. ([112]) used the RGB-D data of human face and SVM classifier was employed to recognize the emotions. The authors in ([223]) presented a method of detecting human emotions with gait data from Kinect. In ([224]), a real-time 3D facial emotion detection system was presented. The system extracted the dynamic face features and used neural networks and SVR (Support Vector Regressors) to analyze the 16 AU (Action Units) and to complete the facial emotion recognition. However, emotion recognition with RGB-D images or with speech is not always accurate, because humans can lie and hide emotional states. Humans can easily express a facial emotion and feel another emotion. Hence, identifying the real emotional state is crucial.

The authors in ([259]) described an IRL with mixed-initiative and how the memory-based human-robot interaction strategies worked in the learning environment. Furthermore, in ([260]) an IRL system was developed to help a curious robot learn skills in the grasping task by using the human speech feedback in the dialog loops. Thomaz et. al ([261]) completed the experiments with a teachable robot system and found that the robot learner with human guidance had a better learning ability than the one without human guidance. However, the researches in the literature do not use the multimodal method, which leads to the limited application scenarios of emotion recognition models.

5.4 Methodology

5.4.1 Fusion of multimodal classiers

In this chapter, we selected the basic emotion recognition model from CNN, HMM, SVM, and RF based on the emotion classification accuracy. The model with the highest accuracy was used as the basic model for building the hybrid model structure. Our hybrid emotion recognition model applied the modified confusion matrix to the decision-level fusion part.

Basic model for hybrid model

Our long-life learning structure aims to fusion the thermal feature and the gait feature with modified confusion matrices because the two types of features have different accuracy for emotion recognition, respectively. At first, we will test the unimodal classifiers and select the unimodal emotion classification model with the highest accuracy as the basic model to build a hybrid fusion model.

As told in *Chapter 4*, based on our offline multimodal database, we have tested HMM with thermal data, HMM model with gait data, and RF model with thermal data, RF model with gait data, CNN model with gait data, SVM model with thermal data, and SVM model with gait data. And training size: validation size: testing size is 12:2:1. In addition, CNN model with gait features has only an offline testing accuracy of 55%. The offline accuracy of HMM, SVM, and RF with gait data are 65%, 65%, and 70%, respectively. SVM and RF with thermal features have the accuracy of 55% and 60%, respectively. Hence, we found that RF had the best emotion recognition performance with the highest accuracy with respect to the other models. So, this paper applies RF as the basic machine learning model for emotion recognition.

Hybrid model for online recognition

Instead of feature level fusion, our hybrid emotion recognition model used the decision-level fusion method with the modified confusion matrix of each modality. The confusion matrices were updated during the interactive robot learning loops, which made us see the unimodal classifier's emotion recognition ability for each emotion in the long-life learning process. The two RFs were used for each single channel in our long-life architecture as the RF gave the best accuracy scores for both thermal and gait data in the offline testing. In addition, if the different basic models got the best accuracy in the both data, feature-level fusion would not be possible in our architecture. That was the reason why we selected the decision-level fusion.

The individual basic models only with gait data and the one only with thermal face data have different recognition abilities for the 4 different emotions, differences that can be seen in the confusion matrix of the emotion classification model. During the online testing, we found that the two individual models show distinct emotion recognition performance with different accuracy in each emotion situation. Hence, the decision-level fusion of the two individual models is necessary to make a better recognition accuracy. We developed a new decision-level fusion method with two basic model (namely RF model) for online emotion recognition and the framework is as shown in Figure 5.1. The integration method is based on the modified confusion matrix and the probability vector of each emotion class.

In the paper ([262]), the decision-level integration method for multimodal emotion recognition uses the accuracy of each emotion in each model. However,the confusion matrix shows the more statistical information of classification performance for different classes than the accuracy. This statistical information from the confusion matrix from each model is useful to integrate multiple models. We applied the confusion matrix information to build up the

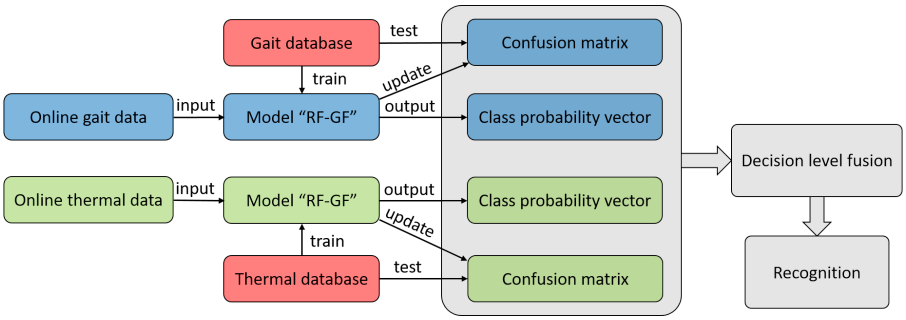


Figure 5.1: The fusion framework of multimodal classifiers

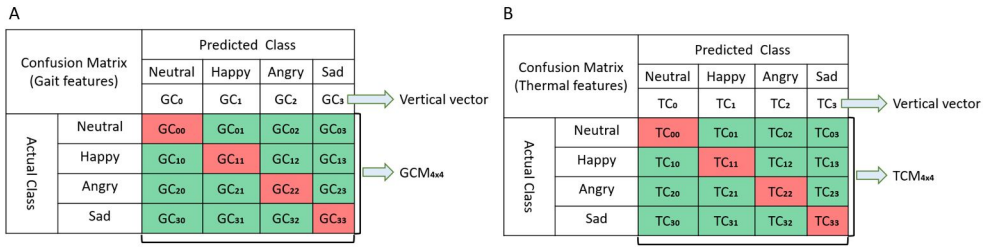


Figure 5.2: Modified confusion matrices. (A) GCM, namely the MCM of the gait model. (B) TCM, namely the MCM of thermal model.

decision-level hybrid model for emotion recognition. In the confusion matrix all the elements represent probabilities. The row of the confusion matrix represents the instances in a predicted class while the column represents the instances in the real class. Then, the elements of each column in the confusion matrix are divided by the total amount of instances in each column, respectively to get the Modified Confusion Matrix (MCM). In MCM, each column shows the probability of each real class in the predicted class situation. MCM of the gait face model and MCM of the thermal model are shown in Figure 5.2 (A) and Figure 5.2 (B), respectively. For example, in the column two corresponding to the predicted class "angry" Figure 5.2 (B), TC_{02}, TC_{12}, TC_{22}, and TC_{32} represent the probabilities of neutral, happy, angry, and sad emotions, respectively when the predicted class is the angry emotion. During every online testing, the class probabilities of the thermal model make up the vector $TPM_{1\times4}$, as shown in Figure 5.3 (B). For example, the thermal emotion recognition model gets the prediction result "happy". From the above matrix, the predicted "happy" is "neutral" with probability TC_{01}, "happy" with probability TC_{11}, "angry" with probability TC_{21}, and "sad" with probability TC_{31}. Therefore, in the "happy" prediction situation, the probability vector of the 4 emotions is equal to $TP_1 \times TC_1$. Similarly, we can get the probability vectors of the 4 emotions in other 3 prediction situations. The related processes for gait features and for thermal features are as shown in Figure 5.3.

The calculation of $TEP_{1\times4}$ for the thermal model, of $GEP_{1\times4}$ for the gait model, of $FUEP_{1\times4}$ for the two new vectors, and of FRR for the final recognition result is as indicated in Equations 5.1, 5.2, 5.3, and 5.4, respectively. The class prediction probability vector of thermal model $TPM_{1\times4}$ times the modified confusion matrix of thermal model $TCM_{4\times4}$ to get the new class probability vector for the thermal modality $TEP_{1\times4}$. Based on the probabilistic history relation among emotions contained in $TCM_{4\times4}$, the prediction probability of thermal model $TPM_{1\times4}$ in online testing is transferred to $TEP_{1\times4}$ with the history information. In the mean time, $GEP_{1\times4}$ is got with the dot product between $GPM_{1\times4}$ and $GCM_{4\times4}$ for the gait modality. Then the addition of $TEP_{1\times4}$ and $GEP_{1\times4}$ is equal to $FUEP_{1\times4}$ which fusions the decision-level information of two modalities. Finally, the hybrid model recognition result is

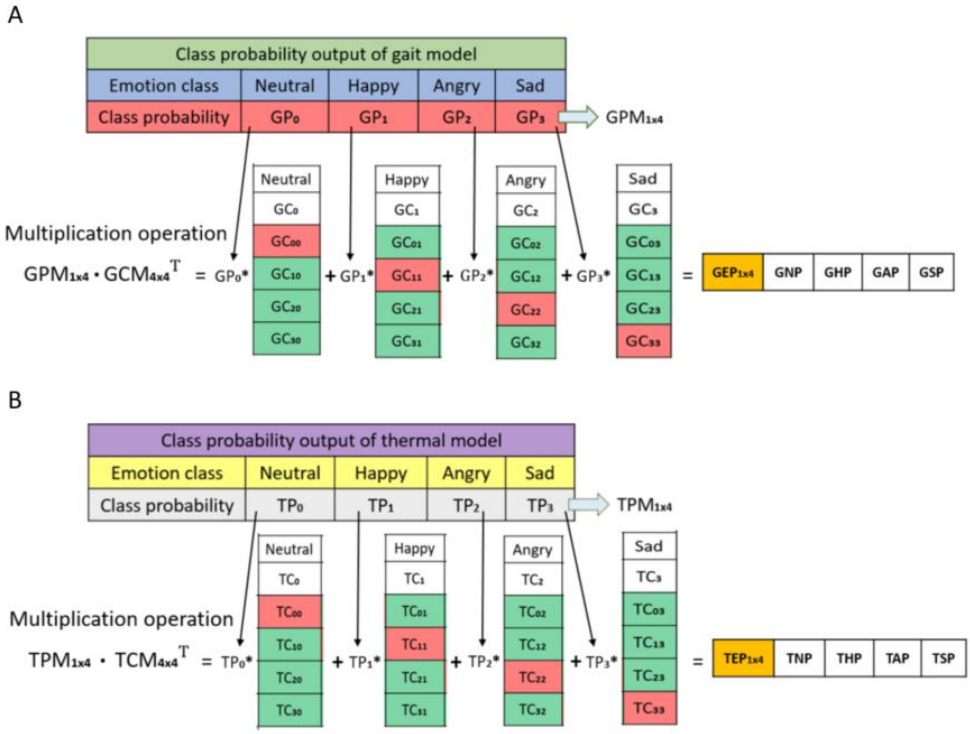


Figure 5.3: (A) Class probability calculation with *MCM* for the gait feature. *GPM* means the class prediction probability vector of gait model and *GCM* means the modified confusion matrix of gait model. Four vertical vectors make of the *GCM*. (B) Class probability calculation with *MCM* for the thermal feature. TPM means the class prediction probability vector of thermal model and TCM means the modified confusion matrix of thermal model. Four vertical vectors make of the TCM.

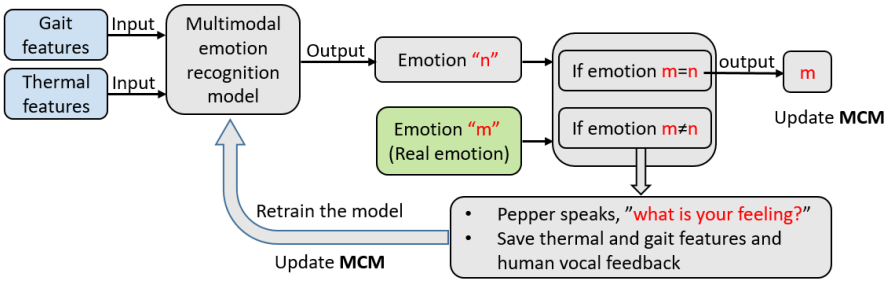


Figure 5.4: Overview of the IRL architecture (emotion "m" and emotion "n" belong to neutral state, happiness, anger, sadness.)

the emotion with the maximum probability in $FUEP_{1\times4}$.

$$TEP_{1\times4} = TPM_{1\times4} \cdot TCM_{4\times4} \tag{5.1}$$

$$GEP_{1\times4} = GPM_{1\times4} \cdot GCM_{4\times4} \tag{5.2}$$

$$FUEP_{1\times4} = TEP_{1\times4} + GEP_{1\times4} \tag{5.3}$$

$$FRR = \text{argmax}\left(FUEP_{1\times4}\right) \tag{5.4}$$

5.4.2 Robot interactive learning model

We applied an interactive learning method with human in the loop during the interaction in order to boost the online emotion recognition performance of the robot. An overview of the IRL system architecture is illustrated in Figure 5.4. During the IRL experiments, gait features and thermal facial features were input into the mulimodal emotion recognition model, which is based on the MCM and is described in section 5.4.1. The recognition result is Emotion "n" (belong to 4 basic emotions including neutral state, happiness, anger and, sadness). If the predicted emotion, namely Emotion "n" does not match the ground truth, namely Emotion "m", the gait and the thermal facial features are restored to retrain the multimodal emotion classification model. The updated thermal and gait basic models in hybrid model test the saved features again to get the new predicted results, respectively. The new predicted results are used to update the two MCMs, which leads to the update of the weight in the decision-level fusion of hybrid model. If the predicted emotion is equal to the real one, we only update the two confusion matrices and the models are not retrained. With humans in the loop again and again, the final updated MCM will show a more precise relation between different emotion labels and finally, the online emotion recognition will be improved.

5.5 Experimental design

5.5.1 Online emotion recognition

We need to consider the computer system issue during online emotion recognition that Kinect V2 with skeleton extraction SDK and the thermal camera with ROS cannot work in the same system. The Kinect V2 cannot work on the Linux system for skeleton-based gait data extraction with Kinect SDK, but the thermal camera with ROS can not work on the Windows system. The related data flow is shown in Figure 5.5. We applied the Kinect on the win10

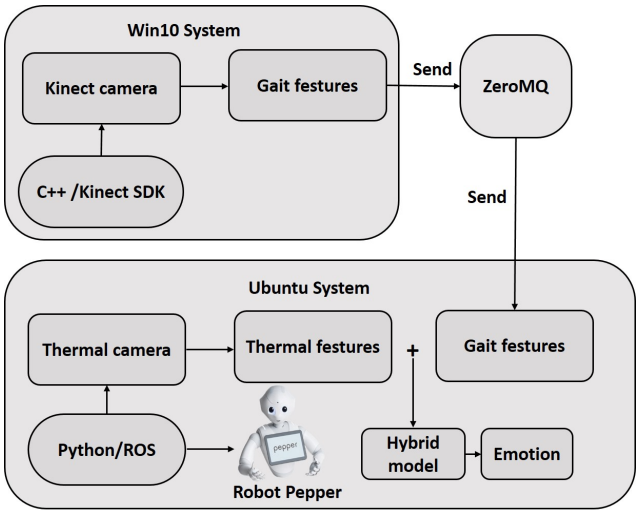


Figure 5.5: Hardware and software architecture.

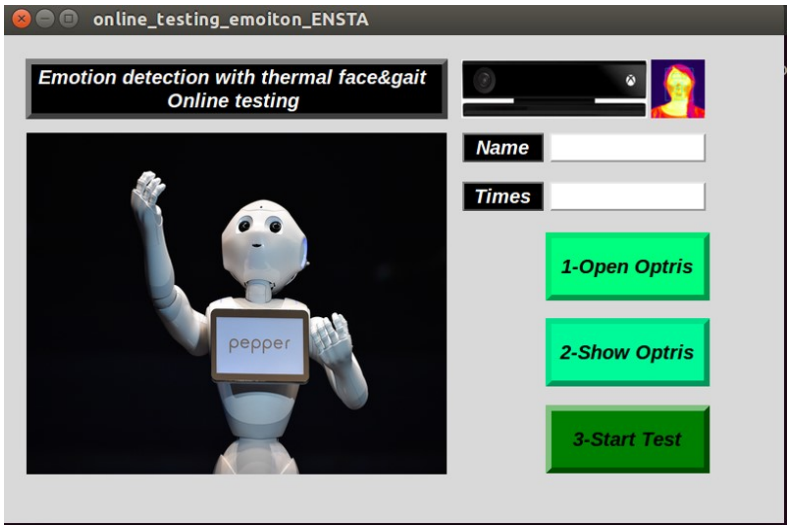


Figure 5.6: Online multimodal emotion recognition interface.

system and the thermal camera on the Ubuntu system. The *ZeroMQ* based message transformation interface [263] was used to send the extracted gait features to the Ubuntu system during the online experiments. The gait features adding the thermal features were fed to the pre-trained hybrid model for online emotion classification. The recognition result will be used for a later interactive robot learning model. To facilitate a convenient interaction, we also built-up an online multimodal emotion recognition interface as shown in Figure 5.6.

5.5.2 Interactive robot learning experiments

The hardware and the environment settings are as depicted in Figure 4.3 and *Chapter 4*. The experiment for the online emotion recognition is composed of three parts: the online testing, the IRL, and the online testing after IRL. 8 participants (half males and half females) participated 8 times in each part. Hence, 192 experiments in total were conducted.

All participants in this research were students with different major backgrounds including Computer Science, Mathematics, Business and others. They came from different universities in Paris with different educational levels from undergraduates to PhD students. All of them could speak English from the intermediate level to the native level and some of them could

speak French as the “FilmStim” for emotion elicitation contains both English and French videos. Their age ranged from 20 to 32. The offline and IRL experiments were conducted with different participants.

In all experiments, the participants joined the emotion elicitation experiments at first. In the literature, static images and films clips are considered as good stimuli to elicit different emotions in the laboratory and films are one of the most effective ways to elicit emotions ([241]). In IRL experiments, the open film clips database-FilmStim ([242]), which is designed for emotion elicitation, was used. The film clips for each emotion were randomly chosen for emotion elicitation. One participant could finish the experiments in several days if he or she could not finish the experiments in one single session. Furthermore, between each emotion elicitation experiment, the participants had a break of about 2-5 minutes. In our experiments, the participants could randomly select videos from “FilmStim”. French participants selected French videos and the others selected English videos. The elicited emotions were compared to the ground truth emotion of “FilmStim” videos and we obtained more than 60%. The effectiveness of the "FilmStim" for emotion elicitation has been certified in the reference [242].

After the emotion elicitation experiments, we applied Pick-A-Mood (PAM) ([264]) as a tool of emotional state measure. As a visual scale, PAM is quick, easy, and reliable. PAM is a character-based pictorial scale for reporting and expressing human moods. It includes three cartoon characters, namely a male, a female, and a robot character. There are nine expressions related to eight distinct emotions and one neutral state in each character. Our research adopted the robot character and selected 4 expressions based on our research. In our paper, the four expression responses were the neutral state, happy emotion, angry emotion, and sad emotion from left to right, as shown in Fig. 5.7. During the emotion measurement test, the participant selected the cartoon character to report his/her mood. In each experiment, the participants needed to fill out the PAM questionnaire two times, namely after the emotion elicitation part and after the interaction with the robot. The result obtained from PAM was used to decide if the data should be recorded. If the two emotion measure results were the same, the data was saved in the database. If not, the data was not saved in the database.

In the multimodal data extraction part of each experiment, there were two stages including the walking stage and the standing stage. Namely, the participant walked towards the robot, then stopped, and stood before the robot to interact for 10 seconds. The details of the experiments (online testing experiments before IRL, IRL experiments, online testing experiments after IRL) are described in the following steps.

(1) The participant randomly selects and watches the film clips from FilmStim for a specific emotion;

(2) The participant selects the emotional state from PAM for mood measurement;

(3) The robot instructs the participant to walk towards it with the specific emotional gait;

(4) The participant walks towards the robot and stops before it for 5 seconds;

(5) The robot asks: "How do you feel?". The robot records the emotion result and the gait/thermal data into a database for retraining the emotion recognition model;

(6) The participant selects the emotional state from PAM for mood measurement again;

(7) Repeat steps from 2 to 6 to complete more experiments.

5.6 Experimental results

5.6.1 Multimodal data analysis

During gait data processing, firstly, the median filter as a non-linear tool was used to remove the noise from the eight gait sequences. After the filtering process, the sequence was

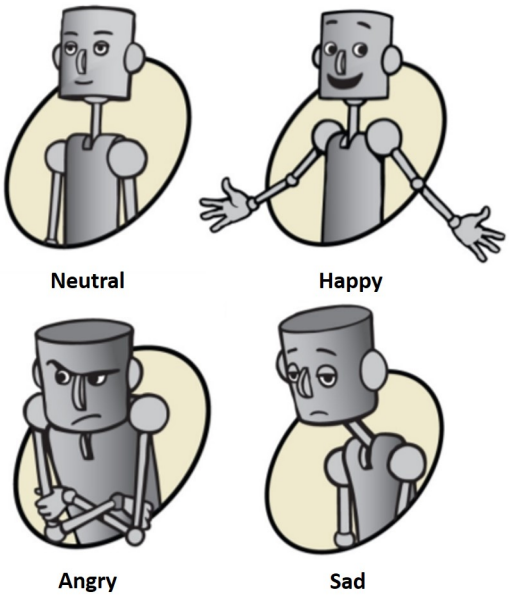


Figure 5.7: Four expression cartoon characters from PAM. The four related emotions are neutral state, happiness, anger and sadness. During the experiment, the emotion labels cannot be seen and the subjects only can see four cartoon characters.

smoothed, which can effectively reduce the reflection of the noise during the machine learning process. Secondly, Welch method ([265]) was used to obtain the PSD features. When using Welch method, 15 frequencies were uniformly selected from 0 Hz to 50 Hz. During thermal feature extraction, the statistic thermography features of face ROIs were got for emotion recognition.

There are 126 multimodal features extracted from the gait and thermal data. Even though our features size is not so large, we investigated the relevance between these features and the emotion labels through Maximal Information coefficient (MIC) method, which is a statistic coefficient to measure the linear or non-linear relation strength for the paired variables [266] and also can be used for feature selection [267]. The MIC results between those features and the labels are shown in Figure 5.8. The MIC values range from 0 to 1 and the larger the value, the stronger the relation between the feature and the label.

5.6.2 Emotion recognition results

Before online testing for emotion recognition, the offline RF models were tested in order to get modified confusion matrix, which is used for fusion of our emotion recognition models. There were 80 testing examples were used for offline testing. The modified confusion matrix shows the probability distribution relation of the four emotions as shown in Table 5.1 and Table 5.2, respectively. In the modified confusion matrix, each column sums up to 100% (if without rounding) instead of each row in the original confusion matrix. And, the primary diagonal elements in the confusion matrix represent accuracy of each emotion while modified confusion matrix's diagonal elements just indicate a conditional probability instead of accuracy of each emotion. In Table 5.2, the ratio with 91.67% is significantly higher when the real class and the predicted class both are sad one. That ratio in the modified confusion matrix only represents the probability of "sad" when predicted class is "sad", instead of the accuracy of "sad" (only 55%) in the confusion matrix.

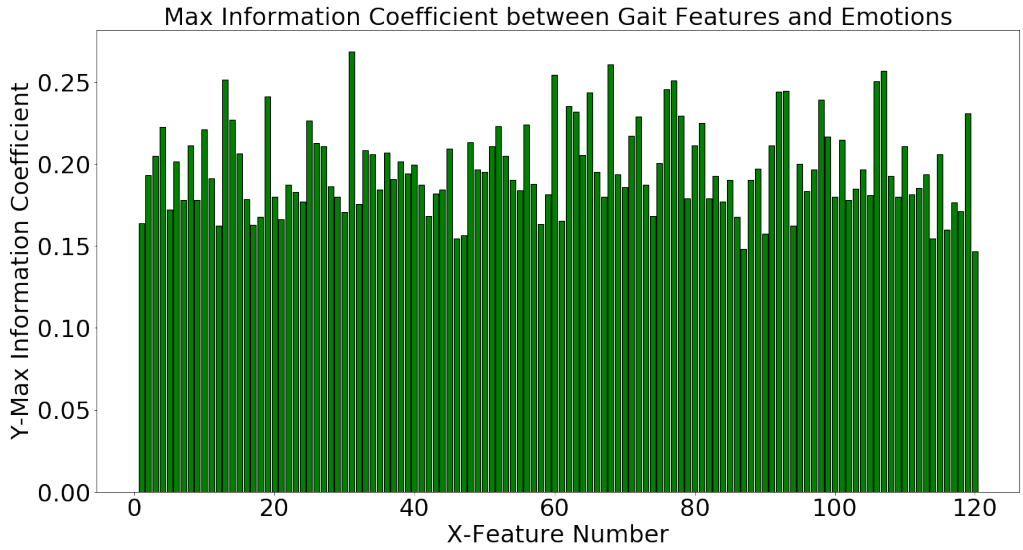


Figure 5.8: Maximal Information coefficient (MIC) between features and labels. X axis is the feature number from 1 to 126. Y axis is related MIC value from 0 to 1. Bigger the MIC value means a stronger correlations between the feature and the label.

Table 5.1: Modified confusion matrix of offline testing of gait model (N: Neutral H: Happy A: Angry S:Sad).

Gait (model)		Predicted class			
		N	H	A	S
Real class	N	75%	16%	0%	6.67%
	H	0%	52%	30%	6.67%
	A	10%	20%	65%	0%
	S	15%	12%	5%	86.67%

Table 5.2: Modified confusion matrix of offline testing of thermal mode (N: Neutral H: Happy A: Angry S:Sad).

Thermal model		Predicted class			
		N	H	A	S
Real class	N	52%	25%	5.26%	0%
	H	12%	50%	21.05%	8.33%
	A	16%	20.83%	57.89%	0%
	S	20%	4.17%	15.79%	91.67%

In the offline part, because the emotional gait data can be acted more easily and the emotional thermal facial data with temperature is hard to be acted, the gait model had a better accuracy for the neutral, and angry emotion than for the thermal model while both models had similar classification capability for the other two emotions. For the integration of the model with gait features and the one with thermal facial features, we applied a decision-level method with the modified confusion matrix. Before IRL experiments, we conducted the online testing experiments with the decision-level model. In the online testing experiments before IRL, the chapter compared the recognition performance of the single models with gait features or thermal facial features and the multimodal model with both, which are with the accuracy of 54.6875% (gait model), 59.375% (thermal model), and 65.625% (fusion model), respectively. The results indicate a higher accuracy for the hybrid model than for any single models.

In order to verify the effectiveness of IRL for long-life emotion recognition, we conducted 3 parts of online experiments, including the online testing, the IRL, and the online testing after IRL. MCM (including TCM and GCM) shows the weight information of fusion part of hybrid model. Before IRL, the weigh information of decision-level fusion can be seen in Table 5.1 and Table 5.2, which was got from the offline testing part. Bigger percent in GCM and TCM will make more contribution in the fusion part for this prediction class. During IRL, the interactive data from human-robot interaction experiments were used to retrain the two basic models in hybrid model and to update the weight information. The updated weight results can be seen in Table 5.3 and Table 5.4. With more interactive data used to improve the emotion recognition structure, more realistic probabilistic relation between the predicted emotions and real emotions were got, which can lead to a better online emotion classification performance in the long-life learning. As shown in Table 5.3 and Table 5.4, the percents were updated compared with the ones before IRL shown in Table 5.1 and Table 5.2.

Table 5.3: Modified confusion matrix of gait model after IRL (N: Neutral H: Happy A: Angry S:Sad).

Gait (model)		Predicted class			
		N	H	A	S
Real class	N	76.32%	12.82%	2.33%	4.17%
	H	2.63%	61.54%	23.26%	4.17%
	A	5.26%	15.38%	65.12%	0%
	S	15.79%	10.26%	9.30%	91.67%

Table 5.4: Modified confusion matrix of thermal model after IRL (N: Neutral H: Happy A: Angry S:Sad).

Thermal (model)		Predicted class			
		N	H	A	S
Real class	N	65.85%	17.07%	5%	0%
	H	7.32%	68.29%	10%	4.55%
	A	9.76%	12.20%	67.5%	0%
	S	17.07%	2.44%	17.5%	95.45%

During IRL with human in the loop, when the hybrid model recognition result does not match the real one, the data and the related labels are saved for model retraining. The confusion matrices are updated with the testing results on the two retrained basic models. After IRL, the online testing was conducted again with the updated hybrid model. At last, an online testing accuracy of 78.125% is obtained after IRL, which is an increase of more than

10% than the one before interactive learning (i.e., 65.625%), which demonstrates that the IRL method is appropriate for emotion recognition with gait and thermal facial data.

5.7 Summary

In this chapter, a hybrid model based on the RF model and the modified confusion matrices of two individual models was developed. By comparing the individual RF models and the decision-level hybrid model, we found out that our integration method is better to classify the emotion during human-robot interaction. 192 experiments were conducted, including three parts, namely online testing experiments before IRL, IRL experiments, and online testing after IRL. In these experiments, we compared the emotion recognition performance before and after IRL. We found out that the interactive robot learning performed better, with a 10% increase of the accuracy. Therefore, the first results show that our IRL architecture with human in the loop is a good solution to be used in robot long life learning in human-robot interaction scenarios. For the future work, we would like to conduct more online emotion recognition experiments in the wild in order to improve the long-life learning model with humans in the loop. Because our multimodal database size was limited and all the experiments were conducted in lab surroundings. The thermal facial data is very hard to be imitated for humans in a lab setting even though there are so many useful emotion elicitation methods. Hence, more natural data should be obtained in real life, which will lead to a more accurate emotion prediction during human-robot interaction.

5.8 Thesis Contributions

In summary, our contributions in this chapter are as follows:

- We built-up a hybrid fusion model based on the RF model and the modified confusion matrices. The model is a decision-level fusion method.
- Using the hybrid model, we conducted online multimodal emotion recognition during human-robot interaction.
- We also introduced an interactive robot learning method with the human in the loop to improve online emotion recognition. The IRL method used the user verbal feedback to relabel the data when the robot cannot recognize user emotion rightly. The relabelled sample can be used to train our multimodal emotion recognition model again for performance improvement. The accuracy shows the effectiveness of our IRL model.

This work has been published at [13].

Chapter 6

Summary

To summarize, we explored how to obtain a natural human-robot interaction from the view of multimodal human emotion perception with the human in the loop and from the perspective of speech-driven robot gesture/face action generation with GAN. Our IRL solution also can inspire some other human behavior perception tasks that are possible with the human in the loop. Our crossmodal mapping model from the acoustic modality to the visual modality and the one-to-many generation architecture also motivate other similar tasks, for example, music-to-dance mapping. The thesis contains four parts: speech-driven robot gesture generation, speaking robot face action synthesis, multimodal human emotion recognition, and interactive robot learning for emotion recognition.

- In the robot gesture generation part, we built up a crossmodal audio-visual database with speech audios and aligned 3D gestures extracted from the YouTube TED video collection. We came up with an one-to-many speech-driven gesture synthesis model based on GAN, which can take one speech as input and generate multiple mapped gestures as the human does. And we also retargeted the generated speaking gestures to robot motor actions for the Pepper robot, which can be used in human-robot interaction. Finally, we evaluate the generated co-speech motions. The results demonstrate the effectiveness of our one-to-many gesture generation model.

- In the robot face action generation part, a GAN-based face action generation model was composed to synthesize the face action from speech audio. And the generated face actions were mapped to the face motor control signal of the face robot Zeno. Finally, we conducted the qualitative evaluation and the quantitative evaluation, and the evaluation with human-joined experiments to assess the generated robot face action compared with ground truth. The results show that our model is the potential for natural human-robot interaction.

- In the multimodal emotion recognition part, we built up a multimodal database with thermal face images and 3D skeletal gait data based on human-robot interaction experiments. The gait features were extracted based on the PSD, and thermal facial features were obtained based on temperature statistics of the face ROI. Then multiple unimodal classifiers were tested on our database. Finally, we built up an easy hybrid model to fusion both kinds of features for better emotion recognition.

- In the IRL part, we built up a new hybrid model based on modified confusion matrices for multimodal emotion recognition with the decision-level fusion. We tested the model for online emotion recognition in real-world human-robot interaction through the real-time multimodal emotion recognition interface built by us. Then we came up with an interactive robot learning architecture with the human in the loop where human verbal feedback was used to improve the online emotion recognition model. Finally, we did online experiments with IRL to validate the effectiveness of our IRL solution.

Chapter 7

Synthèse en français

Pouvoir afficher une interaction naturelle a un impact significatif dans la réussite d'une interaction humain-robot (HRI). Quand nous parlons d'une HRI naturelle, nous faisons référence à la fois à la compréhension du comportement multimodal humain et à la génération de comportements verbaux ou non verbaux du robot. Les humains peuvent naturellement communiquer par le biais du langage et de comportements non verbaux. Par conséquent, un robot doit percevoir et comprendre les comportements humains afin d'être capable de produire un comportement multimodal et naturel qui corresponde au contexte social. Dans cette thèse, nous explorons la compréhension du comportement humain et la génération du comportement du robot pour une HRI naturelle. Cela comprend la reconnaissance multimodale des émotions humaines avec des informations visuelles extraites des cameras RGB-D et thermiques, et la synthèse du comportement non verbal du robot.

La perception des émotions humaines en tant que composante fondamentale de la communication joue un rôle important dans le succès des interactions entre un robot et un humain. La reconnaissance des émotions basée sur les comportements humains multimodaux lors d'une HRI peut aider les robots à comprendre les états des utilisateurs et à produire une interaction sociale naturelle. Dans cette thèse, nousinvestiguons la reconnaissance multimodale des émotions avec des informations thermiques du visage et des données de la marche humaine. Une base de données multimodale contenant des images thermiques du visage et des données de la marche en 3D a été créée grâce aux expériences d'HRI. Nous avons testé les différents classificateurs d'émotions unimodaux (c-à-d, CNN, HMM, forêts aléatoires, SVM) et un classificateur d'émotions hybride pour la reconnaissance des émotions hors ligne. Nous avons également exploré un système de reconnaissance des émotions en ligne avec des capacités limitées dans le cadre de l'HRI en temps réel. L'interaction joue un rôle essentiel dans l'apprentissage des compétences pour une communication naturelle. Pour améliorer notre système de reconnaissance des émotions en ligne, nous avons développé un modèle d'apprentissage robotique interactif (IRL) avec l'humain dans la boucle. Le modèle IRL peut appliquer la rétroaction verbale humaine pour étiqueter ou réétiqueter les données pour améliorer le modèle de reconnaissance des émotions dans une situation d'interaction à long terme. Après avoir utilisé le modèle d'apprentissage interactif du robot, le robot a pu obtenir une meilleure précision de reconnaissance des émotions en temps réel.

Les comportements humains non verbaux tels que les gestes et les expressions faciales se produisent spontanément avec la parole, ce qui conduit à une interaction naturelle et expressive. La génération de gestes et d'expressions faciales par la parole est essentielle pour permettre à un robot social d'exposer des signaux sociaux et de mener une HRI réussie. Cette thèse propose une nouvelle architecture temporelle GAN (Generative Adversarial Network) pour une cartographie un-à-plusieurs de la représentation acoustique de la parole aux gestes correspondants du robot humanoïde. Nous avons également développé une base de données audiovisuelle pour entraîner le modèle de génération de gestes à partir de la parole. La base de données comprend les données audio extraites directement des vidéos et les données des gestes humaines. Notre synthétiseur de gestes peut être appliqué à des robots sociaux avec

des bras. Le résultat de l'évaluation montre l'efficacité de notre modèle génératif pour la génération de gestes de robot à partir de la parole. De plus, nous avons développé un synthétiseur d'expression faciale efficace basé sur GAN. Etant donné un signal audio, une séquence faciale synchrone et réaliste est générée. Nous avons testé cette partie avec le robot Zeno.

Chapter 8

Short CV

8.1 Publications and Patents from 2017

- Chuang Yu, Xiaoxuan HEI, and Adriana TAPUS. ”Talking Robot Face Action Generation for Social Robot.”. IROS 2021. 2021 [under review]
- Chuande Liu, Bingtuan Gao, Chuang Yu and Adriana TAPUS, ”Self-protective motion planning for mobile manipulators in a dynamic door-closing workspace.” Industrial Robot. 2021.
- Chuang Yu and Adriana Tapus. ”Long-life Emotion Recognition for Social Robot with Human in the Loop.” ACM Transactions on Autonomous and Adaptive Systems. [under review]
- Chuang Yu and Adriana Tapus. "SRG3: Speech-driven Robot Gesture Generation with GAN." 2020 16th International Conference on Control Automation Robotics & Vision (ICARCV). 2020.
- Chuang Yu and Adriana Tapus. "Multimodal Emotion Recognition with Thermal and RGB-D Cameras for Human-Robot Interaction." Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. 2020.
- Chuang Yu and Adriana Tapus. "Interactive robot learning for multimodal emotion recognition." International Conference on Social Robotics (ICSR). Springer, Cham, 2019.
- Chuang Yu “Emotion Detection from Gait and Thermal Face for Social Robot.” Journee de L’ ED Interfaces. 2018.
- Chuang Yu and Adriana Tapus. “Human Emotion Detection based on Gait Data and Facial Thermal Images.” Journee-Robotique & I.A. 2018.
- Mei Shuai, Chuang Yu, Mingyue Song. ”A kind of vision-based gait assessment method in the exoskeleton robot system.” CN108022248A, 2018-05-11 [Patent]
- Mei Shuai, Mingyue Song, Chuang Yu. ”A kind of intelligent sensing footwear gait analysis system based on plantar pressure.” CN108013878A, 2018-05-11 [Patent]

8.2 Teaching Assistant

- ROB311 2019-2020: Robotics+AI, IP-Paris, France, Eight classes. Language: English
- ROB311 2020-2021: Robotics+AI, IP-Paris, France, Eight classes. Language: English
- SoBo 2020-2021: Social Robotics, IP-Paris, France, Ten classes. Language: English

8.3 Reviewer for Professional Journals and Conferences

- IROS 2021: International Conference on Intelligent Robots and Systems
- CIS-RAM 2019: 9th IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and Robotics, Automation and Mechatronics (RAM)
- RO-MAN 2019: 28th IEEE International Conference on Robot and Human Interactive Communication
- ICSR 2020: 12th International Conference on Social Robotics
- 2020 International Journal of Social Robotics

8.4 Prizes

- Nov. 2018: Best poster in Journee de L'ED Interfaces (France)

Chapter 9

Appendix A: Big Five questionnaire

Here are the questions of the Big Five questionnaire [268] [215] used during the assessment of robot face action generation. Each participant should answer on a 5-point Likert scale: 1-disagree strongly, 2-disagree a little, 3-neither agree nor disagree, 4-agree a little, and 5-agree strong.

1. Is talkative?
2. Tends to find fault with others?
3. Does a thorough job?
4. Is depressed, blue?
5. Is original, comes up with new ideas?
6. Is reserved?
7. Is helpful and unselfish with others?
8. Can be somewhat careless?
9. Is relaxed, handles stress well?
10. Is curious about many different things?
11. Is full of energy?
12. Starts quarrels with others?
13. Is a reliable worker?
14. Can be tense?
15. Is ingenious, a deep thinker?
16. Generates a lot of enthusiasm?
17. Has a forgiving nature?
18. Tends to be disorganized?
19. Worries a lot?
20. Has an active imagination?
21. Tends to be quiet?
22. Is generally trusting?
23. Tends to be lazy?
24. Is emotionally stable, not easily upset?
25. Is inventive?
26. Has an assertive personality?
27. Can be cold and aloof?
28. Perseveres until the task is finished?
29. Can be moody?
30. Values artistic, aesthetic experiences?
31. Is sometimes shy, inhibited?
32. Is considerate and kind to almost everyone?
33. Does things efficiently?
34. Remains calm in tense situations?
35. Prefers work that is routine?
36. Is outgoing, sociable?

37. Is sometimes rude to others?
38. Makes plans and follows through with them?
39. Gets nervous easily?
40. Likes to reflect, play with ideas?
41. Has few artistic interests?
42. Likes to cooperate with others?
43. Is easily distracted?
44. Is sophisticated in art, music, or literature?
45. Is concerned about privacy issues?

References

[1] Fatemeh Noroozi et al. “Survey on emotional body gesture recognition”. In: *IEEE transactions on affective computing* (2018).

[2] Zhijie Fang, David Vázquez, and Antonio M López. “On-board detection of pedestrian intentions”. In: *Sensors* 17.10 (2017), p. 2193.

[3] Moshe Gabel et al. “Full body gait analysis with Kinect”. In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2012, pp. 1964–1967.

[4] Christiane Goulart et al. “Emotion analysis in children through facial emissivity of infrared thermal imaging”. In: *PloS one* 14.3 (2019), e0212928.

[5] David–Octavian Iacob and Adriana Tapus. “First Attempts in Deception Detection in HRI by using Thermal and RGB-D cameras”. In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2018, pp. 652–658.

[6] Amirhossein H Memar and Ehsan T Esfahani. “Eeg correlates of motor control difficulty in physical human-robot interaction: A frequency domain analysis”. In: *2018 IEEE Haptics Symposium (HAPTICS)*. IEEE. 2018, pp. 229–234.

[7] Maged S Al-Quraishi et al. “EEG-based control for upper and lower limb exoskeletons and prostheses: A systematic review”. In: *Sensors* 18.10 (2018), p. 3342.

[8] Luzheng Bi, Cuntai Guan, et al. “A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration”. In: *Biomedical Signal Processing and Control* 51 (2019), pp. 113–127.

[9] Tatsuya Teramae, Tomoyuki Noda, and Jun Morimoto. “EMG-based model predictive control for physical human–robot interaction: Application for assist-as-needed control”. In: *IEEE Robotics and Automation Letters* 3.1 (2017), pp. 210–217.

[10] Jung-Hoon Kim et al. “Design of a knee exoskeleton using foot pressure and knee torque sensors”. In: *International Journal of Advanced Robotic Systems* 12.8 (2015), p. 112.

[11] Mohamed Yacine Tsalamlal et al. “Affective handshake with a humanoid robot: How do participants perceive and combine its facial and haptic expressions?” In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, pp. 334–340.

[12] Yann LeCun, Yoshua Bengio, et al. “Convolutional networks for images, speech, and time series”. In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[13] Chuang Yu and Adriana Tapus. “Interactive robot learning for multimodal emotion recognition”. In: *International Conference on Social Robotics*. Springer. 2019, pp. 633–642.

[14] Mona Fathollahi Ghezelghieh, Rangachar Kasturi, and Sudeep Sarkar. “Learning camera viewpoint using CNN to improve 3D body pose estimation”. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 685–693.

[15] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[16] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[17] Ruhul Amin Khalil et al. “Speech emotion recognition using deep learning techniques: A review”. In: *IEEE Access* 7 (2019), pp. 117327–117345.

[18] Abbas Saliimi Lokman and Mohamed Ariff Ameedeen. “Modern chatbot systems: A technical review”. In: *Proceedings of the future technologies conference*. Springer. 2018, pp. 1012–1023.

[19] Son Thai Ly et al. “Emotion recognition via body gesture: Deep learning model coupled with keyframe selection”. In: *Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence*. 2018, pp. 27–31.

[20] Yi-Lin Lin and Gang Wei. “Speech emotion recognition based on HMM and SVM”. In: *2005 international conference on machine learning and cybernetics*. Vol. 8. IEEE. 2005, pp. 4898–4901.

[21] Ira Cohen, Ashutosh Garg, Thomas S Huang, et al. “Emotion recognition from facial expressions using multilevel HMM”. In: *Neural information processing systems*. Vol. 2. Citeseer. 2000.

[22] Shibani Hamsa et al. “Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier”. In: *IEEE Access* 8 (2020), pp. 96994–97006.

[23] Shamane Siriwardhana et al. “Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion”. In: *IEEE Access* 8 (2020), pp. 176274–176285.

[24] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *arXiv preprint arXiv:1409.3215* (2014).

[25] Noé Tits et al. “Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis”. In: *arXiv preprint arXiv:1903.11570* (2019).

[26] Yuxuan Wang et al. “Uncovering latent style factors for expressive speech synthesis”. In: *arXiv preprint arXiv:1711.00520* (2017).

[27] Jonathan Shen et al. “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 4779–4783.

[28] Yi Ren et al. “Fastspeech: Fast, robust and controllable text to speech”. In: *arXiv preprint arXiv:1905.09263* (2019).

[29] Ye Jia et al. “Transfer learning from speaker verification to multispeaker text-to-speech synthesis”. In: *arXiv preprint arXiv:1806.04558* (2018).

[30] Ya-Jie Zhang et al. “Learning latent representations for style control and transfer in end-to-end speech synthesis”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6945–6949.

[31] Youngwoo Yoon et al. “Speech gesture generation from the trimodal context of text, audio, and speaker identity”. In: *ACM Transactions on Graphics (TOG)* 39.6 (2020), pp. 1–16.

[32] Carlos T Ishi et al. “A speech-driven hand gesture generation method and evaluation in android robots”. In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 3757–3764.

[33] Maha Salem et al. “Generation and evaluation of communicative robot gesture”. In: *International Journal of Social Robotics* 4.2 (2012), pp. 201–217.

[34] Pierrick Milhorat et al. “A conversational dialogue manager for the humanoid robot ERICA”. In: *Advanced Social Interaction with Agents*. Springer, 2019, pp. 119–131.

[35] Michelle J Salvador, Sophia Silver, and Mohammad H Mahoor. “An emotion recognition comparative study of autistic and typically-developing children using the zeno robot”. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2015, pp. 6128–6133.

[36] Jong Won Kwak et al. “A face robot actuated with artificial muscle based on dielectric elastomer”. In: *Journal of mechanical science and technology* 19.2 (2005), pp. 578–588.

[37] Samer Al Moubayed et al. “Furhat: a back-projected human-like robot head for multiparty human-machine interaction”. In: *Cognitive behavioural systems*. Springer, 2012, pp. 114–130.

[38] Samer AL Moubayed, Gabriel Skantze, and Jonas Beskow. “The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction”. In: *International Journal of Humanoid Robotics* 10.01 (2013), p. 1350005.

[39] Jeong Woo Park, Hui Sung Lee, and Myung Jin Chung. “Generation of realistic robot facial expressions for human robot interaction”. In: *Journal of Intelligent & Robotic Systems* 78.3 (2015), pp. 443–462.

[40] Nikolaos Mavridis. “A review of verbal and non-verbal human–robot interactive communication”. In: *Robotics and Autonomous Systems* 63 (2015), pp. 22–35.

[41] Harish Ravichandar et al. “Recent advances in robot learning from demonstration”. In: *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), pp. 297–330.

[42] Neziha Akalin and Amy Loutfi. “Reinforcement learning approaches in social robotics”. In: *Sensors* 21.4 (2021), p. 1292.

[43] Jordan Abdi et al. “Scoping review on the use of socially assistive robot technology in elderly care”. In: *BMJ open* 8.2 (2018), e018815.

[44] Jan F Veneman et al. “Design and evaluation of the LOPES exoskeleton robot for interactive gait rehabilitation”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 15.3 (2007), pp. 379–386.

[45] Robin R Murphy. “Human-robot interaction in rescue robotics”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34.2 (2004), pp. 138–153.

[46] Alberto Tellaeche, Inaki Maurtua, and Aitor Ibarguren. “Human robot interaction in industrial robotics. Examples from research centers to industry”. In: *2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA)*. IEEE. 2015, pp. 1–6.

[47] Alexander Bannat et al. “A multimodal human-robot-interaction scenario: Working together with an industrial robot”. In: *International conference on human-computer interaction*. Springer. 2009, pp. 303–311.

[48] Valeria Villani et al. “Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications”. In: *Mechatronics* 55 (2018), pp. 248–266.

[49] *What is a Collaborative Robot -COBOT?* https://www.yaskawa-global.com/product/robotics/collaborative. (Accessed on 05/04/2021).

[50] Maria Kyrarini et al. “Robot learning of industrial assembly task via human demonstrations”. In: *Autonomous Robots* 43.1 (2019), pp. 239–257.

[51] Angel Perez Garcia, Ingrid Schjølberg, and Serge Gale. “EEG control of an industrial robot manipulator”. In: *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE. 2013, pp. 39–44.

[52] Wafa Batayneh et al. “Using emg signals to remotely control a 3d industrial robotic arm”. In: *ASME International Mechanical Engineering Congress and Exposition*. Vol. 59414. American Society of Mechanical Engineers. 2019, V004T05A009.

[53] Gilbert Tang and Phil Webb. “The design and evaluation of an ergonomic contactless gesture control system for industrial robots”. In: *Journal of Robotics* 2018 (2018).

[54] Patrik Gustavsson et al. “Human-robot collaboration demonstrator combining speech recognition and haptic control”. In: *Procedia CIRP* 63 (2017), pp. 396–401.

[55] Maria Kyrarini et al. “A Survey of Robots in Healthcare”. In: *Technologies* 9.1 (2021), p. 8.

[56] Mingxing Lyu et al. “Design of a biologically inspired lower limb exoskeleton for human gait rehabilitation”. In: *Review of Scientific Instruments* 87.10 (2016), p. 104301.

[57] Mei Shuai et al. *A kind of vision-based gait assessment method in the exoskeleton robot system*. CN108022248A, May 2018.

[58] Mei Shuai et al. *A kind of intelligent sensing footwear gait analysis system based on plantar pressure*. CN108013878A, May 2018.

[59] Bernhard Specht et al. “Real-Time Robot Reach-To-Grasp Movements Control Via EOG and EMG Signals Decoding”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 3812–3817.

[60] Iroju Olaronke, Ojerinde Oluwaseun, and Ikono Rhoda. “State of the art: a study of human-robot interaction in healthcare”. In: *International Journal of Information Engineering and Electronic Business* 9.3 (2017), p. 43.

[61] Serhan Coşar et al. “ENRICHME: Perception and Interaction of an Assistive Robot for the Elderly at Home”. In: *International Journal of Social Robotics* 12.3 (2020), pp. 779–805.

[62] Roxana Agrigoroaie, François Ferland, and Adriana Tapus. “The enrichme project: Lessons learnt from a first interaction with the elderly”. In: *International Conference on Social Robotics*. Springer. 2016, pp. 735–745.

[63] Claudia Salatino et al. “The enrichme project”. In: *International Conference on Computers Helping People with Special Needs*. Springer. 2016, pp. 326–334.

[64] S Coşar et al. “Thermal camera based physiological monitoring with an assistive robot”. In: *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2018, pp. 5010–5013.

[65] Derek McColl et al. “A survey of autonomous human affect detection methods for social robots engaged in natural HRI”. In: *Journal of Intelligent & Robotic Systems* 82.1 (2016), pp. 101–133.

[66] Beatrice Alenljung et al. “User experience in social human-robot interaction”. In: *Rapid automation: Concepts, methodologies, tools, and applications*. IGI Global, 2019, pp. 1468–1490.

[67] Iolanda Leite, Carlos Martinho, and Ana Paiva. “Social robots for long-term interaction: a survey”. In: *International Journal of Social Robotics* 5.2 (2013), pp. 291–308.

[68] Andrés A Ramırez-Duque et al. “Collaborative and inclusive process with the autism community: a case study in Colombia about social robot design”. In: *International Journal of Social Robotics* (2020), pp. 1–15.

[69] Meia Chita-Tegmark and Matthias Scheutz. “Assistive Robots for the Social Management of Health: A Framework for Robot Design and Human–Robot Interaction Research”. In: *International Journal of Social Robotics* (2020), pp. 1–21.

[70] Anna Furnari et al. “Robotic-assisted gait training in Parkinson’s disease: a three-month follow-up randomized clinical trial”. In: *International journal of neuroscience* 127.11 (2017), pp. 996–1004.

[71] Luthffi Idzhar Ismail et al. “Leveraging robotics research for children with autism: a review”. In: *International Journal of Social Robotics* 11.3 (2019), pp. 389–410.

[72] Jabar H Yousif, Hussein A Kazem, and Miqdam T Chaichan. “Evaluation Implementation of Humanoid Robot for Autistic Children: A Review”. In: *International Journal of Computation and Applied Sciences* 6.1 (2019), pp. 412–420.

[73] Hideki Kozima, Marek P Michalowski, and Cocoro Nakagawa. “Keepon”. In: *International Journal of Social Robotics* 1.1 (2009), pp. 3–18.

[74] Elisabeta Marinoiu et al. “3d human sensing, action and emotion recognition in robot assisted therapy of children with autism”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2158–2167.

[75] Chris Lytridis et al. “Social engagement interaction games between children with Autism and humanoid robot NAO”. In: *The 13th international conference on soft computing models in industrial and environmental applications*. Springer. 2018, pp. 562–570.

[76] Andrew Valenti et al. “Emotion expression in a socially assistive robot for persons with Parkinson’s disease”. In: *Proceedings of the 13th ACM international conference on pervasive technologies related to assistive environments*. 2020, pp. 1–10.

[77] Amanda JC Sharkey. “Should we welcome robot teachers?” In: *Ethics and Information Technology* 18.4 (2016), pp. 283–297.

[78] Omar Mubin et al. “A review of the applicability of robots in education”. In: *Journal of Technology in Education and Learning* 1.209-0015 (2013), p. 13.

[79] Laurent Gallon et al. “Using a Telepresence robot in an educational context”. In: *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*. The Steering Committee of The World Congress in Computer Science, Computer . . . 2019, pp. 16–22.

[80] Jian Liao and Xiaofei Lu. “Exploring the affordances of telepresence robots in foreign language learning”. In: *Language Learning & Technology* 22.3 (2018), pp. 20–32.

[81] Elizabeth Cha, Samantha Chen, and Maja J Mataric. “Designing telepresence robots for K-12 education”. In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 683–688.

[82] Saira Anwar et al. “A systematic review of studies on educational robotics”. In: *Journal of Pre-College Engineering Education Research (J-PEER)* 9.2 (2019), p. 2.

[83] Sofia Serholt. “Child–robot interaction in education”. PhD thesis. University of Gothenburg, 2017.

[84] Yasser Mohammad. “Natural Human-Robot Interaction”. In: *The Wiley Handbook of Human Computer Interaction* 2 (2018), pp. 641–655.

[85] Enrique Coronado et al. “Gesture-based robot control: Design challenges and evaluation with humans”. In: *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2017, pp. 2761–2767.

[86] Stefan Waldherr, Roseli Romero, and Sebastian Thrun. “A gesture based interface for human-robot interaction”. In: *Autonomous Robots* 9.2 (2000), pp. 151–173.

[87] Félix Suárez Bonilla and Federico Ruiz Ugalde. “Automatic translation of spanish natural language commands to control robot comands based on lstm neural network”. In: *2019 third ieee international conference on robotic computing (irc)*. IEEE. 2019, pp. 125–131.

[88] Deng Yongda, Li Fang, and Xin Huang. “Research on multimodal human-robot interaction based on speech and gesture”. In: *Computers & Electrical Engineering* 72 (2018), pp. 443–454.

[89] Matthew P Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. “The right kind of unnatural: designing a robot voice”. In: *Proceedings of the 1st International Conference on Conversational User Interfaces*. 2019, pp. 1–2.

[90] Laurel D Riek et al. “Cooperative gestures: Effective signaling for humanoid robots”. In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2010, pp. 61–68.

[91] Yuichiro Yoshikawa et al. “Responsive robot gaze to interaction partner.” In: *Robotics: Science and systems*. Citeseer. 2006, pp. 37–43.

[92] Pakpoom Patompak et al. “Learning proxemics for personalized human–robot social interaction”. In: *International Journal of Social Robotics* (2019), pp. 1–14.

[93] Stephen W Littlejohn and Karen A Foss. *Encyclopedia of communication theory*. Vol. 1. Sage, 2009, pp. 690–694.

[94] Marc Schröder. “Emotional speech synthesis: A review”. In: *Seventh European Conference on Speech Communication and Technology*. 2001.

[95] Ohsung Kwon et al. “Emotional speech synthesis based on style embedded Tacotron2 framework”. In: *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*. IEEE. 2019, pp. 1–4.

[96] Jesin James et al. “Empathetic Speech Synthesis and Testing for Healthcare Robots”. In: *International Journal of Social Robotics* (2020), pp. 1–19.

[97] Matthew P Aylett, Alessandro Vinciarelli, and Mirjam Wester. “Speech synthesis for the generation of artificial personality”. In: *IEEE transactions on affective computing* 11.2 (2017), pp. 361–372.

[98] Kwan Min Lee et al. “Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction”. In: *Journal of communication* 56.4 (2006), pp. 754–772.

[99] Matthew P Aylett, Yolanda Vazquez-Alvarez, and Skaiste Butkute. "Creating robot personality: effects of mixing speech and semantic free utterances". In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 2020, pp. 110–112.

[100] Clifford Nass, Youngme Moon, and Nancy Green. "Are machines gender neutral? Gender-stereotypic responses to computers with voices". In: *Journal of applied social psychology* 27.10 (1997), pp. 864–876.

[101] Sam Thellman et al. "He is not more persuasive than her: No gender biases toward robots giving speeches". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 2018, pp. 327–328.

[102] Isidoros Rodomagoulakis et al. "Multimodal human action recognition in assistive human-robot interaction". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 2702–2706.

[103] Maha Salem et al. "A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction". In: *2011 Ro-Man*. IEEE. 2011, pp. 247–252.

[104] Iolanda Leite et al. "Long-term interactions with empathic robots: Evaluating perceived support in children". In: *International Conference on Social Robotics*. Springer. 2012, pp. 298–307.

[105] Iolanda Leite et al. "As time goes by: Long-term evaluation of social presence in robotic companions". In: *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2009, pp. 669–674.

[106] Ana Paiva, Iolanda Leite, and Tiago Ribeiro. "Emotion modeling for social robots". In: *The Oxford handbook of affective computing* (2014), pp. 296–308.

[107] James A Coan and John JB Allen. *Handbook of emotion elicitation and assessment*. Oxford university press, 2007.

[108] Georgios Drakopoulos et al. "Emotion Recognition from Speech: A Survey." In: *WEBIST*. 2019, pp. 432–439.

[109] Javier G Rázuri et al. "Speech emotion recognition in emotional feedbackfor human-robot interaction". In: *International Journal of Advanced Research in Artificial Intelligence (IJARAI)* 4.2 (2015), pp. 20–27.

[110] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. "Multimodal speech emotion recognition using audio and text". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2018, pp. 112–118.

[111] Alejandro Lopez-Rincon. "Emotion recognition using facial expressions in children using the NAO Robot". In: *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*. IEEE. 2019, pp. 146–153.

[112] Qi-rong Mao et al. "Using Kinect for real-time emotion recognition via facial expressions". In: *Frontiers of Information Technology & Electronic Engineering* 16.4 (2015), pp. 272–282.

[113] Anushree Basu et al. "Human emotion recognition from facial thermal image based on fused statistical feature and multi-class SVM". In: *2015 Annual IEEE India Conference (INDICON)*. IEEE. 2015, pp. 1–5.

[114] Daniel McDuff et al. "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit". In: *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 2016, pp. 3723–3726.

[115] Richard Savery, Ryan Rose, and Gil Weinberg. "Establishing human-robot trust through music-driven robotic emotion prosody and gesture". In: *2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE. 2019, pp. 1–7.

[116] Christiana Tsiourti et al. "Designing emotionally expressive robots: a comparative study on the perception of communication modalities". In: *Proceedings of the 5th international conference on human agent interaction*. 2017, pp. 213–222.

[117] Antonio Chella et al. "An Emotional Storyteller Robot." In: *AAAI spring symposium: emotion, personality, and social behavior*. 2008, pp. 17–22.

[118] Myrthe Tielman et al. "Adaptive emotional expression in robot-child interaction". In: *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2014, pp. 407–414.

[119] Jaime A Rincon et al. "A new emotional robot assistant that facilitates human interaction and persuasion". In: *Knowledge and Information Systems* 60.1 (2019), pp. 363–383.

[120] Michael Argyle. *Bodily communication*. Routledge, 2013, pp. 1–5.

[121] Henny Admoni et al. "Robot nonverbal behavior improves task performance in difficult collaborations". In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2016, pp. 51–58.

[122] Sébastien Saint-Aimé et al. "EmotiRob: companion robot project". In: *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2007, pp. 919–924.

[123] Carlos T Ishi. "Motion Generation during Vocalized Emotional Expressions and Evaluation in Android Robots". In: *Becoming Human with Humanoid-From Physical Interaction to Social Intelligence*. IntechOpen, 2019.

[124] Jan Ondras et al. "Audio-driven robot upper-body motion synthesis". In: *IEEE transactions on cybernetics* (2020).

[125] Marynel Vázquez et al. "Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze". In: *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE. 2017, pp. 42–52.

[126] Pierre-Henri Orefice et al. "Pressure variation study in human-human and human-robot handshakes: Impact of the mood". In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2018, pp. 247–254.

[127] S Craig Roberts, Jan Havlıček, and Benoist Schaal. *Human olfactory communication: current challenges and future prospects*. 2020.

[128] Dylan F Glas et al. "Erica: The erato intelligent conversational android". In: *2016 25th IEEE International symposium on robot and human interactive communication (RO-MAN)*. IEEE. 2016, pp. 22–29.

[129] Alisa Kalegina et al. "Characterizing the design space of rendered robot faces". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 2018, pp. 96–104.

[130] Yunkyung Kim and Bilge Mutlu. "How social distance shapes human–robot interaction". In: *International Journal of Human-Computer Studies* 72.12 (2014), pp. 783–795.

[131] Aniket Bera et al. “Sociosense: Robot navigation amongst pedestrians with social and psychological constraints”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 7018–7025.

[132] Amir Aly and Adriana Tapus. “Multimodal adapted robot behavior synthesis within a narrative human-robot interaction”. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, pp. 2986–2993.

[133] Adriana Tapus, Mataric Maja, and Brian Scassellatti. “The grand challenges in socially assistive robotics”. In: *IEEE Robotics and Automation Magazine* 14.1 (2007), N–A.

[134] Jane Vincent et al. *Social robots from a human perspective*. Springer, 2015, pp. 1–10.

[135] Haibin Yan, Marcelo H Ang, and Aun Neow Poo. “A survey on perception methods for human–robot interaction in social robots”. In: *International Journal of Social Robotics* 6.1 (2014), pp. 85–119.

[136] Adriana Tapus et al. “Children with autism social engagement in interaction with Nao, an imitative robot: A series of single case experiments”. In: *Interaction studies* 13.3 (2012), pp. 315–347.

[137] Selma Šabanović et al. “PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia”. In: *2013 IEEE 13th international conference on rehabilitation robotics (ICORR)*. IEEE. 2013, pp. 1–6.

[138] Luke J Wood et al. “Developing kaspar: a humanoid robot for children with autism”. In: *International Journal of Social Robotics* (2019), pp. 1–18.

[139] Elizabeth S Kim et al. “Social robots as embedded reinforcers of social behavior in children with autism”. In: *Journal of autism and developmental disorders* 43.5 (2013), pp. 1038–1049.

[140] Meritxell Valentı Soler et al. “Social robots in advanced dementia”. In: *Frontiers in aging neuroscience* 7 (2015), p. 133.

[141] Adriana Tapus et al. “Perceiving the person and their interactions with the others for social robotics–a review”. In: *Pattern Recognition Letters* 118 (2019), pp. 3–13.

[142] Chuang Yu and Adriana Tapus. “Multimodal Emotion Recognition with Thermal and RGB-D Cameras for Human-Robot Interaction”. In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 2020, pp. 532–534.

[143] Silvia Rossi, Francois Ferland, and Adriana Tapus. “User profiling and behavioral adaptation for HRI: a survey”. In: *Pattern Recognition Letters* 99 (2017), pp. 3–12.

[144] Amir Aly and Adriana Tapus. “Prosody-based adaptive metaphoric head and arm gestures synthesis in human robot interaction”. In: *2013 16th International Conference on Advanced Robotics (ICAR)*. IEEE. 2013, pp. 1–8.

[145] Patrik Jonell et al. “Learning Non-verbal Behavior for a Social Robot from YouTube Videos”. In: *ICDL-EpiRob Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions, Oslo, Norway, August 19, 2019*. 2019.

[146] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[147] Marc’Aurelio Ranzato et al. “On deep generative models with applications to recognition”. In: *CVPR 2011*. IEEE. 2011, pp. 2857–2864.

[148] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[149] Tero Karras et al. “Audio-driven facial animation by joint end-to-end learning of pose and emotion”. In: *ACM Transactions on Graphics (TOG)* 36.4 (2017), pp. 1–12.

[150] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. “End-to-end speech-driven facial animation with temporal gans”. In: *arXiv preprint arXiv:1805.09313* (2018).

[151] Hsin-Ying Lee et al. “Dancing to Music”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 3581–3591.

[152] Taoran Tang, Jia Jia, and Hanyang Mao. “Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis”. In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 1598–1606.

[153] Weihao Xia et al. “GAN inversion: A survey”. In: *arXiv preprint arXiv:2101.05278* (2021).

[154] Tero Karras et al. “Progressive growing of gans for improved quality, stability, and variation”. In: *arXiv preprint arXiv:1710.10196* (2017).

[155] Zhaoqing Pan et al. “Recent progress on generative adversarial networks (GANs): A survey”. In: *IEEE Access* 7 (2019), pp. 36322–36333.

[156] KR Prajwal et al. “A lip sync expert is all you need for speech to lip generation in the wild”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 484–492.

[157] Eli Shlizerman et al. “Audio to body dynamics”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7574–7583.

[158] Kirsten Bergmann and Stefan Kopp. “GNetIc–Using bayesian decision networks for iconic gesture generation”. In: *International Workshop on Intelligent Virtual Agents*. Springer. 2009, pp. 76–89.

[159] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. “Predicting co-verbal gestures: a deep and temporal modeling approach”. In: *International Conference on Intelligent Virtual Agents*. Springer. 2015, pp. 152–166.

[160] Michael Studdert-Kennedy. “Hand and Mind: What Gestures Reveal About Thought.” In: *Language and Speech* 37.2 (1994), pp. 203–209.

[161] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. “A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents”. In: *arXiv preprint arXiv:2101.03769* (2021).

[162] Paul Bremner et al. “Beat gesture generation rules for human-robot interaction”. In: *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2009, pp. 1029–1034.

[163] Paul Bremner and Ute Leonards. “Iconic gestures for robot avatars, recognition and integration with speech”. In: *Frontiers in psychology* 7 (2016), p. 183.

[164] Justine Cassell et al. “Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents”. In: *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 1994, pp. 413–420.

[165] Amir Aly and Adriana Tapus. “Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction”. In: *Autonomous Robots* 40.2 (2016), pp. 193–209.

[166] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. "Beat: the behavior expression animation toolkit". In: *Life-Like Characters*. Springer, 2004, pp. 163–185.

[167] Youngwoo Yoon et al. "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 4303–4309.

[168] Dai Hasegawa et al. "Evaluation of speech-to-gesture generation using bi-directional LSTM network". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 2018, pp. 79–86.

[169] Taras Kucherenko et al. "Analyzing input and output representations for speech-driven gesture generation". In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 2019, pp. 97–104.

[170] Simon Alexanderson et al. "Generating coherent spontaneous speech and gesture from text". In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 2020, pp. 1–3.

[171] Simon Alexanderson et al. "Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows". In: *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, pp. 487–496.

[172] Shiry Ginosar et al. "Learning individual styles of conversational gesture". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3497–3506.

[173] Kurima Sakai et al. "Novel Speech Motion Generation by Modeling Dynamics of Human Speech Production". In: *Frontiers in Robotics and AI* 4 (2017), p. 49.

[174] Taras Kucherenko et al. "Gesticulator: A framework for semantically-aware speech-driven gesture generation". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020, pp. 242–250.

[175] Hao Zhu et al. "Deep audio-visual learning: A survey". In: *International Journal of Automation and Computing* (2021), pp. 1–26.

[176] Dimitri Palaz, Ronan Collobert, et al. *Analysis of cnn-based speech recognition system using raw speech as input*. Tech. rep. Idiap, 2015.

[177] Wei Dai et al. "Very deep convolutional neural networks for raw waveforms". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 421–425.

[178] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. 1. 2013, p. 3.

[179] Triantafyllos Kefalas et al. "Speech-driven facial animation using polynomial fusion of features". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 3487–3491.

[180] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[181] Sven Franz, Ralph Nolte-Holube, and Frank Wallhoff. "NAFOME: NAO Follows Me-tracking, reproduction and simulation of human motion". In: *Jade University of Applied Sciences, Germany* (2013).

[182] Emrehan Yavşan and Ayşegül Uçar. "Gesture imitation and recognition using Kinect sensor and extreme learning machines". In: *Measurement* 94 (2016), pp. 852–861.

[183] *Pepper-Joints-Aldebaran 2.0.6.8 documentation*. (Date last accessed 20th-May-2020). URL: http://doc.aldebaran.com/2-0/family/juliette_technical/joints%5C_juliette.html.

[184] Tae-Hyun Oh et al. "Speech2face: Learning the face behind a voice". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7539–7548.

[185] Zhe Cao et al. "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7291–7299.

[186] Bellard Fabrice. *FFmpeg*. Version 4.1.5. Jan. 7, 2020. URL: https://ffmpeg.org.

[187] Julieta Martinez et al. "A simple yet effective baseline for 3d human pose estimation". In: *ICCV*. 2017.

[188] Catalin Ionescu et al. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (July 2014), pp. 1325–1339.

[189] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[190] Chuang Yu and Adriana Tapus. "SRG 3: Speech-driven Robot Gesture Generation with GAN". In: *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE. 2020, pp. 759–766.

[191] Youngwoo Yoon et al. "Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity". In: *ACM Transactions on Graphics* 39.6 (2020).

[192] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. "End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 37–40.

[193] Ikhsanul Habibie et al. "A recurrent variational autoencoder for human motion synthesis". In: *28th British Machine Vision Conference*. 2017.

[194] Jina Lee and Stacy Marsella. "Nonverbal behavior generator for embodied conversational agents". In: *International Workshop on Intelligent Virtual Agents*. Springer. 2006, pp. 243–255.

[195] David Foster. *Generative deep learning: teaching machines to paint, write, compose, and play*. O'Reilly Media, 2019.

[196] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation". In: *arXiv preprint arXiv:1703.10847* (2017).

[197] Agrim Gupta et al. "Social gan: Socially acceptable trajectories with generative adversarial networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2255–2264.

[198] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *Advances in neural information processing systems*. 2017, pp. 5767–5777.

[199] Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: *International Conference on Machine Learning*. 2015, pp. 1530–1538.

[200] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. "Moglow: Probabilistic and controllable motion synthesis using normalising flows". In: *arXiv preprint arXiv:1905.06598* (2019).

[201] Volker Blanz and Thomas Vetter. "A morphable model for the synthesis of 3D faces". In: *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 1999, pp. 187–194.

[202] Bernhard Egger et al. "3D Morphable Face Models—Past, Present, and Future". In: *ACM Transactions on Graphics (TOG)* 39.5 (2020), pp. 1–38.

[203] Yang Zhou et al. "MakeItTalk: Speaker-Aware Talking Head Animation". In: *arXiv preprint arXiv:2004.12992* (2020).

[204] Najmeh Sadoughi and Carlos Busso. "Speech-driven expressive talking lips with conditional sequential generative adversarial networks". In: *IEEE Transactions on Affective Computing* (2019).

[205] Ahmed Hussen Abdelaziz et al. "Modality Dropout for Improved Performance-driven Talking Faces". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020, pp. 378–386.

[206] Carlos Toshinori Ishi, Takashi Minato, and Hiroshi Ishiguro. "Analysis and generation of laughter motions, and evaluation in an android robot". In: *APSIPA Transactions on Signal and Information Processing* 8 (2019).

[207] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 2017, pp. 214–223.

[208] Mohammad Shayganfar, Charles Rich, and Candace L Sidner. "A design methodology for expressing emotion on robot faces". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2012, pp. 4577–4583.

[209] Gabriele Fanelli et al. "A 3-d audio-visual corpus of affective communication". In: *IEEE Transactions on Multimedia* 12.6 (2010), pp. 591–598.

[210] Md Sahidullah and Goutam Saha. "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". In: *Speech communication* 54.4 (2012), pp. 543–565.

[211] Namrata Dave. "Feature extraction methods LPC, PLP and MFCC in speech recognition". In: *International journal for advance research in engineering and technology* 1.6 (2013), pp. 1–4.

[212] Joseph W Picone. "Signal modeling techniques in speech recognition". In: *Proceedings of the IEEE* 81.9 (1993), pp. 1215–1247.

[213] Davis E King. "Dlib-ml: A machine learning toolkit". In: *The Journal of Machine Learning Research* 10 (2009), pp. 1755–1758.

[214] *Open3D library*. http://www.open3d.org/docs/release/getting_started.html. (Accessed on 05/23/2021).

[215] Oliver P John, Laura P Naumann, and Christopher J Soto. "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues." In: (2008).

[216] VR Preedy and RR Watson. "5-point Likert scale". In: *Handbook of Disease Burdens and Quality of Life Measures, Watson RR (ed): Springer, New York* (2010).

[217] Chuang Yu, Xiaoxuan Hei, and Adriana Tapus. "Robot Facial Action Generation from Speech with GAN". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE [Under Review].

[218] A. Kale, A.K. RoyChowdhury, and R. Chellappa. “Fusion of gait and face for human identification”. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 5. IEEE. 2004, pp. 899–901.

[219] E. Hossain and G. Chetty. “Multimodal identity verification based on learning face and gait cues”. In: *International Conference on Neural Information Processing*. Springer. 2011, pp. 1–8.

[220] X. Zhou and B. Bhanu. “Feature fusion of side face and gait for video-based human identification”. In: *Pattern Recognition* 41.3 (2008), pp. 778–795.

[221] X. Geng et al. “Context-aware fusion: A case study on fusion of gait and face for human identification in video”. In: *Pattern recognition* 43.10 (2010), pp. 3660–3673.

[222] G. Castellano, S. Villalba, and A. Camurri. “Recognising Human Emotions from Body Movement and Gesture Dynamics”. In: *Affective Computing and Intelligent Interaction* (2007), pp. 71–82.

[223] S. Li et al. “Emotion recognition using Kinect motion capture data of human gaits”. In: *PeerJ* 4 (2016), e2364.

[224] Y. Zhang, L. Zhang, and M. Hossain. “Adaptive 3D facial action intensity estimation and emotion recognition”. In: *Expert Systems with Applications* 42.3 (2015), pp. 1446–1464.

[225] C.L. Roether et al. “Critical features for the perception of emotion from gait”. In: *Journal of vision* 9.6 (2009), pp. 15–15.

[226] M.M. Khan, R.D. Ward, and M. Ingleby. “Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature”. In: *ACM Transactions on Applied Perception (TAP)* 6.1 (2009), p. 6.

[227] Michelle Karg, Kolja Kühnlenz, and Martin Buss. “Recognition of affect based on gait patterns”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.4 (2010), pp. 1050–1061.

[228] Arthur Crenn et al. “Body expression recognition from animated 3D skeleton”. In: *2016 International Conference on 3D Imaging (IC3D)*. IEEE. 2016, pp. 1–7.

[229] Uttaran Bhattacharya et al. “Step: Spatial temporal graph convolutional networks for emotion perception from gaits”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 02. 2020, pp. 1342–1350.

[230] Sijie Yan, Yuanjun Xiong, and Dahua Lin. “Spatial temporal graph convolutional networks for skeleton-based action recognition”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[231] Venkatraman Narayanan et al. “Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation”. In: *arXiv preprint arXiv:2003.01062* (2020).

[232] L. Boccanfuso et al. “A thermal emotion classifier for improved human-robot interaction”. In: *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2016, pp. 718–723.

[233] Y. Yoshitomi et al. “Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face”. In: *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499)*. IEEE. 2000, pp. 178–183.

[234] R. Nakanishi and K. Imai-Matsumura. "Facial skin temperature decreases in infants with joyful expression". In: *Infant Behavior and Development* 31.1 (2008), pp. 137–144.

[235] Z. Wu, M. Peng, and T. Chen. "Thermal face recognition using convolutional neural network". In: *Optoelectronics and Image Processing (ICOIP), 2016 International Conference on*. IEEE. 2016, pp. 6–9.

[236] L. Trujillo et al. "Automatic feature localization in thermal images for facial expression recognition". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE. 2005, pp. 14–14.

[237] Mustafa MM Al Qudah, Ahmad SA Mohamed, and Syaheerah L Lutfi. "Affective State Recognition Using Thermal-Based Imaging: A Survey". In: *COMPUTER SYSTEMS SCIENCE AND ENGINEERING* 37.1 (2021), pp. 47–62.

[238] I.A. Cruz-Albarran et al. "Human emotions detection based on a smart-thermal system of thermographic images". In: *Infrared Physics & Technology* 81 (2017), pp. 250–261.

[239] Shangfei Wang et al. "Fusion of visible and thermal images for facial expression recognition". In: *Frontiers of Computer Science* 8.2 (2014), pp. 232–242.

[240] Meike K Uhrig et al. "Emotion elicitation: A comparison of pictures and films". In: *Frontiers in psychology* 7 (2016), p. 180.

[241] Yaling Deng, Meng Yang, and Renlai Zhou. "A new standardized emotional film database for Asian culture". In: *Frontiers in psychology* 8 (2017), p. 1941.

[242] A. Schaefer et al. "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers". In: *Cognition and Emotion* 24.7 (2010), pp. 1153–1172.

[243] S. Wang et al. "A natural visible and infrared facial expression database for expression recognition and emotion inference". In: *IEEE Transactions on Multimedia* 12.7 (2010), pp. 682–691.

[244] M. Karg et al. "A comparison of PCA, KPCA and LDA for feature extraction to recognize affect in gait kinematics". In: *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*. IEEE. 2009, pp. 1–6.

[245] I.S. MacKenzie. "Within-subjects vs. Between-subjects Designs: Which to Use?" In: *Human-Computer Interaction: An Empirical Research Perspective* 7 (2002), p. 2005.

[246] Rawesak Tanawongsuwan and Aaron Bobick. "Gait recognition from time-normalized joint-angle trajectories in the walking plane". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 2. IEEE. 2001, pp. II–II.

[247] P. Welch. "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms". In: *IEEE Transactions on audio and electroacoustics* 15.2 (1967), pp. 70–73.

[248] P.K. Rahi and R. Mehra. "Analysis of power spectrum estimation using welch method for various window techniques". In: *International Journal of Emerging Technologies and Engineering* 2.6 (2014), pp. 106–109.

[249] K. Barbe, R. Pintelon, and J. Schoukens. "Welch method revisited: nonparametric power spectrum estimation via circular overlap". In: *IEEE Transactions on signal processing* 58.2 (2010), pp. 553–565.

[250] R.B. Blackman and J.W. Tukey. “The measurement of power spectra from the point of view of communications engineering—Part I”. In: *Bell System Technical Journal* 37.1 (1958), pp. 185–282.

[251] E. Salazar-López et al. “The mental and subjective skin: Emotion, empathy, feelings and thermography”. In: *Consciousness and cognition* 34 (2015), pp. 149–162.

[252] Y. Sugimoto, Y. Yoshitomi, and S. Tomita. “A method for detecting transitions of emotional states using a thermal facial image based on a synthesis of facial expressions”. In: *Robotics and Autonomous Systems* 31.3 (2000), pp. 147–160.

[253] R. Agrigoroaie, A. Cruz-Maya, and A. Tapus. “”Oh! I am so sorry!”: Understanding User Physiological Variation while Spoiling a Game Task”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 313–319.

[254] J.B. Yang et al. “Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition”. In: *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000*. Citeseer, 2015, pp. 3995–4001.

[255] M. Brand, N. Oliver, and A. Pentland. “Coupled hidden Markov models for complex action recognition”. In: *Computer vision and pattern recognition, 1997. proceedings., 1997 ieee computer society conference on*. IEEE. 1997, pp. 994–999.

[256] L.R. Brody. “On understanding gender differences in the expression of emotion”. In: *Human feelings: Explorations in affect development and meaning* (1993), pp. 87–121.

[257] T.M. Chaplin. “Gender and emotion expression: A developmental contextual perspective”. In: *Emotion Review* 7.1 (2015), pp. 14–21.

[258] D. Katagami and S. Yamada. “Interactive classifier system for real robot learning”. In: *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499)*. IEEE. 2000, pp. 258–263.

[259] M. Hanheide and G. Sagerer. “Active memory-based interaction strategies for learning-enabling behaviors”. In: *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2008, pp. 101–106.

[260] I. Lutkebohle et al. “The curious robot-structuring interactive robot learning”. In: *2009 IEEE International Conference on Robotics and Automation*. IEEE. 2009, pp. 4156–4162.

[261] A.L. Thomaz and C. Breazeal. “Teachable robots: Understanding human teaching behavior to build more effective robot learners”. In: *Artificial Intelligence* 172.6-7 (2008), pp. 716–737.

[262] L.A. Perez-Gaspar, S.O. Caballero-Morales, and F. Trujillo-Romero. “Multimodal emotion recognition with evolutionary computation for human-robot interaction”. In: *Expert Systems with Applications* 66 (2016), pp. 42–61.

[263] Ravi Prakash Joshi. *Kinect_anywhere: Kinect v2 for ROS while using Kinect Windows API*. https://github.com/ravijo/kinect_anywhere. (Accessed on 05/27/2021).

[264] P. MA Desmet, M.H. Vastenburg, and N. Romero. “Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale”. In: *Journal of Design Research* 14.3 (2016), pp. 241–279.

[265] P. Welch. "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms". In: *IEEE Transactions on audio and electroacoustics* 15.2 (1967), pp. 70–73.

[266] David N Reshef et al. "Detecting novel associations in large data sets". In: *science* 334.6062 (2011), pp. 1518–1524.

[267] Chen Lin et al. "Maximal information coefficient for feature selection for clinical document classification". In: *ICML Workshop on Machine Learning for Clinical Data. Edingburgh, UK*. 2012.

[268] *The Big Five Personality Test*. https://onlineprivacyfoundation.org/Big5PE/quiz.php. (Accessed on 05/24/2021).

ECOLE
DOCTORALE

Titre : Génération du Comportement du Robot et Compréhension du Comportement Humain dans L'interaction Naturelle Homme-Robot

Mots clés : interaction naturelle homme-robot, reconnaissance multimodale des émotions, génération de gestes de robot, génération d'action de visage de robot, apprentissage interactif de robot

Résumé : Dans cette thèse, nous explorons la compréhension du comportement humain et la génération du comportement du robot pour une HRI naturelle. Cela comprend la reconnaissance multimodale des émotions humaines avec des informations visuelles extraites des cameras RGB-D et thermiques, et la synthèse du comportement non verbal du robot.
La perception des émotions humaines en tant que composante fondamentale de la communication joue un rôle important dans le succès des interactions entre un robot et un humain. La reconnaissance des émotions basée sur les comportements humains multimodaux lors d'une HRI peut aider les robots à comprendre les états des utilisateurs et à produire une interaction sociale naturelle. Dans cette thèse, nousinvestiguons la reconnaissance multimodale des émotions avec des informations thermiques du visage et des données de la marche humaine. Pour améliorer notre système de reconnaissance des émotions en ligne, nous avons développé un modèle d'apprentissage robotique interactif (IRL) avec l'humain dans la boucle. Le modèle IRL peut appliquer la rétroaction verbale humaine pour étiqueter ou réétiqueter les données pour améliorer le modèle de reconnaissance des émotions dans une situation d'interaction à long terme. Après avoir utilisé le modèle d'apprentissage interactif du robot, le robot a pu obtenir une meilleure précision de reconnaissance des émotions en temps réel.
Les comportements humains non verbaux tels que les gestes et les expressions faciales se produisent spontanément avec la parole, ce qui conduit à une interaction naturelle et expressive. La génération de gestes et d'expressions faciales par la parole est essentielle pour permettre à un robot social d'exposer des signaux sociaux et de mener une HRI réussie. Cette thèse propose une nouvelle architecture temporelle GAN (Generative Adversarial Network) pour une cartographie un-à-plusieurs de la représentation acoustique de la parole aux gestes correspondants du robot humanoïde. De plus, nous avons développé un synthétiseur d'expression faciale efficace basé sur GAN. Etant donné un signal audio, une séquence faciale synchrone et réaliste est générée. Nous avons testé cette partie avec le robot Zeno.

Title : Robot Behavior Generation and Human Behavior Understanding in Natural Human-Robot Interaction

Keywords : Natural human-robot interaction, multimodal emotion recognition, robot gesture generation, robot face action generation, interactive robot learning

Abstract : In this thesis, we explore human behavior understanding and robot behavior generation for natural HRI. This includes multimodal human emotion recognition with visual information extracted from RGB-D and thermal cameras and non-verbal multimodal robot behavior synthesis.
Emotion recognition based on multimodal human behaviors during HRI can help robots understand user states and exhibit a natural social interaction. In this thesis, we explored multimodal emotion recognition with thermal facial information and 3D gait data in HRI scene when the emotion cues from thermal face and gait data are difficult to disguise. To improve our online emotion recognition system, we developed an interactive robot learning (IRL) model with the human in the loop. The IRL model can apply the human verbal feedback to label or relabel the data for retraining the emotion recognition model in a long-term interaction situation. After using the interactive robot learning model, the robot could obtain a better emotion recognition accuracy in real-time HRI.
The human non-verbal behaviors such as gestures and face action occur spontaneously with speech, which leads to a natural and expressive interaction. Speech-driven gesture and face action generation are vital to enable a social robot to exhibit social cues and conduct a successful HRI. This thesis proposes a new temporal GAN (Generative Adversarial Network) architecture for a one-to-many mapping from acoustic speech representation to the humanoid robot's corresponding gestures. Moreover, we developed an effective speech-driven facial action synthesizer based on GAN, i.e., given an acoustic speech, a synchronous and realistic 3D facial action sequence is generated for face robot Zeno.

Institut Polytechnique de Paris
91120 Palaiseau, France