DEVELOPMENT OF A ROBOTIC NANNY FOR CHILDREN AND A CASE STUDY OF EMOTION RECOGNITION IN HUMAN-ROBOTIC INTERACTION

Yan Haibin
(B.Eng, M.Eng, XAUT)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE
2012

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

闫海滨

Yan Haibin

7 August 2012

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisors, Prof. Marcelo H Ang Jr and Prof. Poo Aun-Neow. Their enthusiastic supervision and invaluable guidance have been essential for the results presented here. I am very grateful that they have spent much time with me to discuss different research problems. Their knowledge, suggestions, and discussions help me to become a more capable researcher. Their encouragement also helps me to overcome the difficulties encountered in my research.

I would also like to express my thanks to other members in our group, Dai Dongjiao Tiffany, Dev Chandan Behera, Cheng Chin Yong, Wang Niyou, Shen Zihong, Kwek Choon Sen Alan, and Lim Hewei, who were involved to help the development of our robot Dorothy Robotubby.

In addition, I would like to thank Mr. Wong Hong Yee Alvin from A*STAR, I^2R and Prof. John-John Cabibihan from National University of Singapore for their valuable suggestions and comments that have helped us to design a pilot study to evaluate our developed robot.

Next, I would like to thank Prof. Marcelo H Ang Jr, Prof. John-John Cabibihan, Mrs. Tuo Yingchong, Mrs. Zhao Meijun, and their family members who were

involved in our pilot studies to evaluate the developed robot.

Lastly, my sincere thanks to Department of Mechanical Engineering, National University of Singapore, Singapore, for providing the full research scholarship to me to support my Ph.D study.

Table of Contents

Summary

With the rapid development of current society, parents become more busy and cannot always stay with their children. Hence, a robotic nanny which can care for and play with children is desirable. A robotic nanny is a class of social robots acting as a child's caregiver and aims to extend the length of parents or caregiver absences by providing entertainment to the child, tutoring the child, keeping the child from physical harm, and ideally, building a companionship with the child. While many social robotics have been developed for children in entertainment, healthcare, and domestic areas, and some promising performance have been demonstrated in their target environments, they cannot be directly applied as a robotic nanny, or cannot satisfy our specific design objectives. Therefore, we develop our own robotic nanny by taking the existing robots as references.

Considering our specific design objectives, we design a robotic nanny named Dorothy Robotubby with a caricatured appearance, which consists of a head, a neck, a body, two arms, two hands, and a touch screen in its belly. Then, we develop two main user interfaces which are local control-based and remote control-based for the child and parents, respectively. Local control-based interface is developed

for a child to control the robot directly to execute some tasks such as telling a story, playing music and games, chatting, and video calling. Remote control-based interface is designed for parents to control the robot remotely to execute several commands like demonstrating facial expressions and gestures when communicating with a child via "video-chat" (like Skype). Since emotion recognition can make important contributions towards achieving a believable and acceptable robot and has become a necessary and significant function in social robotics for a child, we also study facial expression-based emotion recognition by addressing two problems which are important to drive facial expression recognition into real-world applications: misalignment-robust facial expression recognition and cross-dataset facial expression recognition. For misalignment-robust facial expression recognition, we first propose a biased discriminative learning method by imposing large penalties on interclass samples with small differences and small penalties on those samples with large differences simultaneously such that more discriminative features can be extracted for recognition. Then, we learn a robust feature subspace by using the IMage Euclidean Distance (IMED) rather than the widely used Euclidean distance such that the subspace sought is more discriminative and robust to spatial misalignments. For cross-dataset facial expression recognition, we propose a new transfer subspace learning approach to learn a feature space which transfers the knowledge gained from the training set to the target (testing) data to improve the recognition performance under cross-dataset scenarios. Following this idea, we formulate four new transfer subspace learning methods, i.e., transfer principal component analysis (TPCA), transfer linear discriminant analysis (TLDA),

transfer locality preserving projections (TLPP), and transfer orthogonal neighborhood preserving projections (TONPP). Lastly, we design a pilot study to evaluate whether the children like the appearance and functions of Dorothy Robotubby and collect the parents' opinions to the remote user interface designs. To analyze the performance of Robotubby and the interaction between the child and the robot, we employ questionnaires and videotapes. Correspondingly, evaluation results are obtained by questionnaire analysis, behavior analysis, and case studies.

In summary, for misalignment-robust and cross-dataset facial expression recognitions, experimental results have demonstrated the efficacy of our proposed methods. While for the design of our robot Dorothy Robotubby, evaluation results from pilot studies have shown that while there is some room to improve our robotic nanny, most children and parents show great interest in our robot and provide comparatively positive evaluation. More important, several valuable and helpful suggestions are obtained from the result analysis phase.

List of Tables

List of Figures

Chapter 1

Introduction

Social robotics, an important branch of robotics, has recently attracted increasing interest in many disciplines, such as computer vision, artificial intelligence, and mechatronics, and has also emerged as an interdisciplinary undertaking. While many social robots have been developed, a formal definition of social robot has not been agreed on and different practitioners have defined it from different perspectives. For example, Breazeal et al. [1] explained that a social robot is a robot which is able to communicate with humans in a personal way; Fong et al. [2] defined social robots as being able to recognize each other and engage in social interactions; Bartneck and Forlizzi [3] described a social robot as an autonomous or semi-autonomous robot that interacts with humans by following some social behaviors; Hegel et al. [4] defined that a social robot is a combination of a robot and a social interface. In Wikipedia, a social robot [5] is specified to be an autonomous robot that interacts and communicates with humans or other autonomous physical agents by following some social rules. While there are some differences among

these definitions, they have a common characteristic which is to interact with humans. While a great deal of challenges are encountered when social robots are used in real-world applications, there are already some social robots being developed or commercially available to assist our daily lives. They have been used for testing, assisting, and interacting [2]. Depending on their application objects, they can be utilized for the child, the elderly, and the adult.

Among these applications, we mainly focus on developing social robotics for the child in this work. The developed social robotics can not only be used at home to be a child's companion, nanny, for entertainment, but also in several public places like schools, hospitals, and care houses to accomplish some assisting tasks. The robotic companion and nanny can play with and care for the child at home during the absence of busy working parents. Compared with televisions and videos, the robot enables to extend the length of parents' absence. In addition, it can keep the child safe from harm via its monitoring function for a longer time [6]. In public places like hospitals, kindergartens, and care houses, the robots can implement pre-specified tasks to assist nurses and teachers, and can be employed for animal-assisted therapy (AAT) and animal-assisted activities (AAA) instead of real animals [2]. This can partly reduce working strength of the staff, activate learning interest of the child, comfort the child in hospitalization, and provide better therapy to the child with disabilities such as autism [7].

In this study, we aim to develop a robotic nanny to be used at home to take care of a child, play with a child, and activate a child's interest to learn new knowledge. With the rapid development of current society and increasing living pressure, the parents may be very busy and cannot always stay with their children. Under

such situation, a robotic nanny can care for and play with the children during parents' absence. This can release the pressure of parents to a certain extent. Furthermore, due to the concentration of high technologies in the robot, it may activate the child's interest to play with the robot and learn new knowledge during their interaction. The robotic nanny also serves as a two-way communication device with video and physical interaction since the parent can remotely move the limbs of the robotic nanny when interacting with the child.

In the following sections of this chapter, the design objectives of our robotic nanny is introduced. Then, an important emotion recognition function of our robotic nanny is discussed.

1.1 Development of A Robotic Nanny for Children

A robotic nanny is a subclass of social robots which functions as a child's caregiver [8] and aims to extend the length of parent or caregiver absences by providing entertainment to the child, tutoring the child, keeping the child from physical harm, and building a companionship with the child [9, 6]. To develop a satisfactory robotic nanny for children, several design issues related to appearances, functions, and interaction interfaces should be considered [10, 1]. These design problems have a close connection with the application areas and objects of the robot. Generally, different application areas and objects require distinct appearances, functions, and interaction interfaces designs of the robot. For example, the design of a robotic nanny for a child with autism is different from that for

a normal child. In addition to health condition, a child's age, individual difference, personality, and cultural background also play important roles in designing a robotic nanny [8].

AIBO for entertainment, Probo for healthcare, and PaPeRo for childcare are three representative social robotics for a child. While not all of them are designed to be a robotic nanny, their appearances and functions could give us some hints when we develop our own robot for a child.

AIBO is developed by Sony Corporation and is commercially available. From 1999 to 2006, 5 series of this kind of robot were developed [11]. All AIBO series have a dog-like appearance and size, and can demonstrate dog-like behaviors. AIBO is designed to be a robotic companion/pet such that it is autonomous and can learn like a living dog by exploring its world. To behave like a real dog, AIBO has some abilities such as face and object detection and recognition, spoken command recognition, voice detection and recognition, and touch sensing through cameras, microphones, and tactile sensors [12].

Probo, an intelligent huggable robot, is developed to comfort and emotionally interact with the children in a hospital. It has the appearance of an imaginary animal based on ancient mammoths, is about 80cm in height, and moves mainly depending on its fully actuated head [13]. Remarkable features of Probo are its moving trunk and the soft jacket. Due to the soft jacket, the children can make a physical contact with Probo. In addition, Probo has a tele-interface with a touch screen mounted on its belly and a robotic user interface in an external computer. Specifically, the tele-interface is used for entertainment, communication, and

medical assistance, and the robotic user interface is applied to manually control the robot. Probo can also track the ball, detect face and hands, and recognize children's emotional states [14].

PaPeRo is a personal robot designed by the NEC Corporation and commercially available. It can care for children and provide assistance to elders. PaPeRo is about 40cm in height, and has 5 different colors including red, orange, yellow, green, and blue. Unlike the high mobilities of AIBO's body and Probo's head, PaPeRo can only move its head and walk via its wheels [15]. Several application scenarios are developed to make PaPeRo to interact with children, including conversation through speech, face memory and recognition, touching reaction, roll-call and quiz game designing, contacting through phone or PC, learning greetings, and storytelling [16]. Moreover, speakers and LEDs are mounted to produce speech and songs and display PaPeRo's internal status, respectively.

For the above reviewed social robots, it can be seen that AIBO and PaPeRo are commercially available and have been successfully utilized in some real applications such as entertainment and childcare. AIBO can behave like a real pet dog and develop its own unique personality during experiencing its world. Moreover, it can be a research platform for further study. For example, Jones and Deeming [17] proposed an acoustic emotion recognition method and combined it into Sony AIBO ERS7-M3. Since AIBO only behaves like a pet dog, it can only be used in animal pets related applications, which largely limits its application areas. For PaPeRo, it can well execute its predefined scenarios by combining several basic functions such as speech recognition and face tracking. However, it has less mobility as it can only move its head and walk through the wheels. Due to the

less mobility, several functions such as showing the robot's emotions and dancing with more gestures are difficult to be developed.

Different from AIBO and PaPeRo, Probo is not commercially available and is still being developed. Moreover, it has a bigger size such that a touch screen can be mounted on its belly. This is a more direct way to fulfill child-robot interaction. Based on the touch screen, functions like video playing can be included. In addition, another interface used to manually control the robot has been developed in Probo such that the robot becomes an intermedium between the operator and the child, which is especially useful for the child with autism. However, similar to PaPeRo, Probo also has less mobility as it only has a fully actuated head. It is difficult to make Probo to demonstrate more gestures, which may reduce the child's interest.

Since different social robots have their own target environments, there are large differences among their appearances, functions, and interaction interfaces designs. Consequentially, it is difficult to simultaneously use the current developed social robots for a child in different application areas due to their distinct design objectives. Therefore, the researchers should develop their own robot if the existing social robots cannot satisfy their requirements.

Based on the review of the above robots, it can be seen that they cannot be directly applied as a robotic nanny, or cannot satisfy our design objectives. They can only be used as references. The specific design gaps in relation to these robots are summarized as below:

(1) For appearance design, while the above reviewed robots have appealing appearances to a child, some of them are unsuitable for a robotic nanny, such as AIBO. AIBO is designed as a pet dog [12], and it may be difficult to let a child accept a pet dog as his/her nanny. Therefore, to design a robotic nanny with an acceptable appearance should be considered.

(2) Function design has a closer relationship with application areas and objects compared with appearance design. In addition, it depends largely on appearance design. Since our robotic nanny has the specific application area and the unique appearance design, the functions of other robots cannot be directly applied for our robot like storytelling of PaPeRo [16] and video playing of Probo [14] due to their different representation forms and contents. Moreover, several new functions should be developed to characterize our own robotic nanny.

(3) For the interface design, since it is decided by appearance and function designs, it requires more design independence. Such design of other robots can only give some hints such as the interaction interface's layout, color, and operability. According to the appearance and functions of our robotic nanny, it is important to design an interaction interface with good appearances and convenient operability.

In this study, we aim to develop a robotic nanny to play with and take care of a child during his/her parent or caregiver absences. We expect our developed robot can not only interact with a child in an attractive way, but also build a connection between a child and his/her parent. The developed robotic nanny will be used at home and focuses mainly on a normal child.

To satisfy these requirements, we have the following specific objectives:

(1) a robot with a upper body and a caricatured appearance by following Mori's "uncanny valley" [18]. It mainly consists of a head, a neck, a body, two arms, two hands, and a touch screen in its belly.

(2) a robot with several functions by adopting a user-centered design approach [19]. These functions include storytelling, playing music, games, chatting, face tracking, video call, emotion recognition, and remote control.

(3) a robot containing two interaction interfaces in accordance with a user-centered design approach [19]. Specifically, one interface is used to operate the robot by a child, and the other interface is utilized to remotely control the robot by parents.

In addition to developing an acceptable robotic nanny, a real pilot study is designed to evaluate the performance of our developed robot and explore the interaction between the child and the robot. We expect that such a pilot study can be used to improve the current functions and develop new functions of the robot, which makes our robot more fascinating for potential use in other applications.

We expect our robot Dorothy Robotubby is a new member of robotic nannies in the near future. Dorothy Robotubby is the first of a family of social robots with "family name" Robotubby. It may better activate a child's interest to interact with the robot and extend the length of parent or caregiver absences. It can also build a connection between a child and his/her parent. Moreover, it can give several hints to other robotic researchers when they develop their own robots. Our robot will be tested in real pilot studies with children. The testing results will be useful to study child-robot interaction which is significant in children-related topics such as studying child development and providing therapy for disabled children.

In this study, the appearance, function, and interaction interface designs of our robotic nanny are introduced. We mainly concentrate on function and interface designs, especially for the software development part. As for appearance design, it is very complicated and involves several engineering issues like a robot's morphology, mechanical, and electrical designs. These problems are not central to this study and not discussed in detail.

1.2 Emotion Recognition in the Robotic Nanny

As Dautenhahn, Bond, Canamero, and Edmonds [20] stated: "Agents that can recognize a user's emotions, display meaningful emotional expressions, and behave in ways that are perceived as coherent, intentional, responsive, and socially/emotionally appropriate, can make important contributions towards achieving human-computer interaction that is more 'natural', believable, and enjoyable to the human partner." In addition, emotion plays an important role in long-term physical well-being, physiological reactions, cognitive processes, and behavior of humans, especially for children who are in development [8]. Therefore, emotion recognition has become a necessary and significant function in lots of social robots for a child, such as Probo. It senses the user's emotion states by using facial expression and speech [14].

To recognize users' emotion states, there are several cues to be utilized. Generally, these cues can be extracted from visual signals, audio signals, tactile signals, and other channels. For visual signals, facial expression, body language and posture are widely used. They are important for humans to express their emotions.

Specifically, facial expressions can well express humans' emotions including happiness, sadness, fear, anger, disgust, and surprise regardless of culture [21], and body languages and postures are effective cues when facial features are unavailable or unreliable under certain conditions such as at a long distance [22]. These vision-based cues are easily collected with various resolutions, however, they are sensitive to varying illuminations.

For audio signals, speech is a promising way to detect emotions, where emotional information is conveyed by linguistic messages and paralinguistic features [23]. Due to different culture backgrounds, paralinguistic messages like prosody [24] and nonlinguistic vocalizations [23] are more exploited compared with linguistic messages. Similar to visual signals, audio signals are also easily collected. Furthermore, they are low-cost, nonintrusive, and have fast time resolution. However, they are easily affected by the environment noises.

Physical reactions such as touching are usual behaviors during human-human interaction or human-robot interaction. The collected tactile signals contain the emotional content and hence become another useful modal to sense emotions [25]. Different from visual and audio signals, tactile signals are more robust to the varying environments. However, they are heavily influenced by tactile sensors. The type, number, accuracy, mounting places and ways of tactile sensors may affect the final recognition results. Moreover, it is difficult to accurately connect physical reactions with emotional states.

Besides the above modals, other signals representing physiological activities are also employed to recognize emotion. These signals are recordings of electrical

signals produced by muscles, skin, heart, and brain [23]. They usually reflect spontaneous emotions of humans. However, it needs external equipments to collect these signals.

By comparing the advantages and disadvantages of the above used signals and motivated by the fact that most information (∼75%) received for human beings are visual signals, we choose visual signals to recognize the user's emotions. Facial expression, body language and posture are three popular visual signals for emotion recognition. Mehrabian [26] has shown that in human face-to-face communication, only 7% and 38% information are transferred by spoken language and paralanguage, respectively, and 55% is transferred by facial expressions. Based on this reason, we select facial expression to recognize emotions in this study.

1.2.1 Facial Expression-Based Emotion Recognition

Automatic facial expression recognition plays an important role in human emotion perception and social interaction, and has attracted much attention in the areas of pattern recognition, computer vision, human-computer interaction, and human-robot interaction.

Over the past three decades, a number of facial expression analysis methods have been proposed, and they can be mainly classified into two categories: geometry-based and appearance-based. Geometry-based methods usually extract facial features such as the shapes and locations of facial components (like the mouth, eyes, brows and nose) and represent them by a feature vector to characterize the facial geometry [27, 28]. In general, different facial expressions have different feature

representations. Appearance-based methods holistically convert each facial image into a feature vector and then apply subspace analysis techniques to extract some statistical features for facial expression representation [29, 30]. In this study, we apply appearance-based methods for facial expression recognition. This is because it is challenging to precisely localize and extract stable geometrical features such as landmarks in each facial image for geometry-based methods in many practical applications, especially when face images are collected under uncontrolled environments. Moreover, geometry-based methods ignore facial texture information in the extracted features. However, texture information has been widely used in many face analysis tasks such as face recognition and facial expression recognition, and the performance of this feature is reasonably good.

Subspace analysis techniques are representative appearance-based methods and have been widely used to reveal the intrinsic structure of data and applied for facial expression recognition. By using these methods, facial expression images are projected into a low-dimensional feature space to reduce the feature dimensions. Representative such methods include principal component analysis (PCA) [31], linear discriminant analysis (LDA) [32], locality preserving projections (LPP) [33] and orthogonal neighborhood preserving projections (ONPP) [34]. Experimental results on several benchmark face databases have also shown the advantage of this kind of methods.

However, these methods have only demonstrated good performance under their experimental conditions, and shown poor performance under real applications. The specific gaps of existing facial expression recognition methods are summarized below.

(1) Most existing appearance-based facial expression recognition methods can only work well when face images are well-aligned. However, in many real world applications such as human-robot interaction and visual surveillance, it is very challenging to obtain well-aligned face images for recognition, especially under uncontrolled conditions. Hence, there are usually some spatial misalignments in the cropped face images due to the eye localization errors even if the eye positions are manually located. A natural question is how spatial misalignments affect the performance of these appearance-based facial expression recognition methods and how to address this problem if spatial misalignments affect the performance of these appearance-based methods.

(2) Most existing facial expression recognition methods assume facial images in the training and testing sets are collected under the same condition such that they are independent and identically distributed. However, in many real world applications, this assumption may not hold as the testing data are usually collected online and generally more uncontrollable than the training data, such as different races, illuminations and imaging conditions. Under this scenario, the performance of conventional subspace learning methods may be poor because the training and testing data are not independent and identically distributed. The generalization capability of these methods is limited on the cross-dataset facial expression recognition problem.

In this study, we aim to address these two problems that are important to drive facial expression recognition into real-world applications by proposing the following two methods:

(1) a biased linear discriminant analysis (BLDA) method with the IMage Euclidean Distance (IMED) to extract discriminative features for misalignment-robust facial expression recognition.

(2) a new transfer subspace learning approach to improve the performance of cross-dataset facial expression recognition.

By using our proposed methods, the performance of facial expression recognition under uncontrolled scenarios can be improved such that facial expression recognition can be used in several real-world applications such as human-robot interaction.

1.3 Summary

In summary, we mainly aim to achieve the following goals in this thesis.

(1) To develop a robotic nanny that can play with and take care of a child. It will be designed from three aspects: appearance, function, and interaction interface designs.

(2) To propose several advanced machine learning methods to address misalignment-robust facial expression recognition and cross-dataset facial expression recognition.

(3) To design a real pilot study to evaluate the performance of our developed robot and explore the interaction between the child and the robot.

The thesis is organized as follows. Chapter 2 provides a general literature review of representative social robotics for a child and facial expression-based emotion

recognition. Chapter 3 introduces the developed robotic nanny, Dorothy Robotubby. In Chapters 4-5, we study misalignment-robust and cross-dataset facial expression recognitions. Chapter 6 analyzes experimental results by applying the developed robotic nanny in real pilot studies with children. Finally, conclusions and future work are presented in Chapter 7.

Chapter 2

Literature Review

Over the past three decades, a large number of social robotics have been developed for children in the entertainment, healthcare, education, and domestic areas [2]. While some of them are not particularly designed as a robotic nanny, their appearance and function designs could provide us some hints when we develop our own robot for a child. In this chapter, we will review some popular design approaches and issues for building effective social robots and introduce several representative social robotics for a child. Due to the important role of emotion recognition in social robotics for a child, we also briefly review several representative facial expression-based emotion recognition algorithms in this chapter.

2.1 Design A Social Robot for Children

A social robot is an undertaking from multi-disciplines such as mechanical and electrical designs, artificial intelligence, computer vision, control theory, and natural and social sciences. With the rapid development of these disciplines, more and

more social robots have been applied to assist people's daily life. For example, social robots for children have been used in the entertainment, healthcare, childcare, education, and therapy areas. Since many factors such as target environment, gender and age information, cultural and social background, and health status affect the design of social robots, proper design approaches and issues should be considered to successfully develop an acceptable social robot.

2.1.1 Design Approaches and Issues

From a design perspective, Fong et al. [2] classified design approaches into two categories: biologically inspired-based and functionally designed-based. Biologically inspired methods aim to create robots to simulate or mimic living creatures' social behavior and intelligence. This kind of methods generally takes natural and social sciences as theory basis and requires the developed robots to be "life-like". AIBO [12], a robot dog, is a representative example. Functionally designed-based approaches aim to design a socially intelligent robot without following any science or nature theory. They are usually driven by beliefs and desires and focus mainly on the function and performance designs of a robot. The functionally designed robots do not need to have the "life-like" capability. PaPeRo [16], used for childcare, is a representative example.

Having selected a suitable design approach, several design issues should be taken into account. Embodiment is one important factor. Dautenhahn et al. [35] defined that embodiment is "establishing a basis for structural coupling by creating the potential for mutual perturbation between system and environment." Different embodied forms and structures of a robot cause different responses from

the environment. Fong et al. [2] classified social robots' aesthetic forms into four categories: anthropomorphic, zoomorphic, caricatured, and functional.

Anthropomorphic robots, which follow human characteristics, have been widely applied as research platforms to study some scientific theories such as ethology, theory of mind, and development psychology [36]. Humanoid robots are representative examples in this category [37]. This kind of robots is able to support meaningful social interactions due to their high degree of human-likeness. Hence, when designing such robots, it requires to consider the robots' structural and functional appropriateness with people [38].

Zoomorphic robots are developed to imitate living creatures. Specifically, animal counterparts are general embodied forms. Generally, it is easier to design social interaction skills for zoomorphic robots than anthropomorphic robots. That is because human-creature relationships between zoomorphic robots and humans are simpler than human-human relationships between anthropomorphic robots and humans [2]. Most of entertainment robots, personal robots, and toy robots belong to this category.

Caricatured robots are designed in virtual forms instead of realistic livings and agents. This kind of robots normally has specific attributes and can easily give an expressive impression to the users. Due to such specific features, more functions to draw and maintain attention can be developed. Additionally, caricatured robots are capable of providing unusual and uncommon appearances, they are easy to establish a lower social expectation and effectively fulfill intended and biased interactions [10, 38].

For functional robots, they are built according to their objectives and functions. Robots with different applications generally have different forms and structures. This kind of robots focuses on the accomplishment of their functions, and thus the embodiment of functional robots reflects the designed tasks. Service robots are examples of this category [2, 10].

While most existing social robots can be classified into the above four groups, there are some overlaps between the first three categories and the last category. This is due to the fact that the robots belonging to the first three categories also require to accomplish several predefined functions, and it is unavoidable to add some functional features into the robots for their operational objectives. For example, some toy robots with animal appearances belong to zoomorphic robots. However, due to some factors such as the limited production cost, the ability to attract children, and the adaptive capability to various situations, the design of these toy robots should reflect functional requirements. From this perspective, these robots can also be classified into functional robot category [2].

From the above analysis, we find that anthropomorphic and zoomorphic robots follow biologically inspired-based methods and caricatured and functional robots adhere to functionally designed methods. Therefore, when designing a social robot, once the robot's embodied form is determined, the corresponding design approach could be selected. For the embodiment of a robot, it is mainly based on the robot's design objectives. Design objectives can provide lots of useful and important information, such as where the robot is used; who the users are; what the robot executes; and what the robot achieves. According to these information, the used embodiment of a robot can be decided. Correspondingly, the robot's

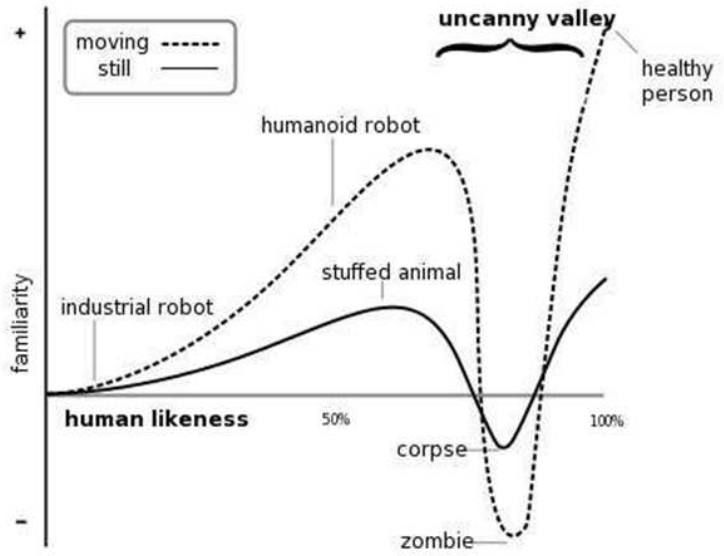


Figure 2.1: The uncanny valley [18].

appearance, functions, and interaction ways can be determined. It is to be noted that these three items should be closely related to design objectives and match each other such that the user can feel natural and comfortable when operating or interacting with the robot.

In addition to the above mentioned design approaches and issues, there is another design theory–Mori's "uncanny valley" hypothesis [18]–to follow. The hypothesis holds that when robots or other human replicas look and act as humans, it causes a response of revulsion among human observers. It is shown in Figure 2.1. Based on this theory, we need to carefully consider how to build anthropomorphic robots. If there is no specific requirement for the developed robot, the other three embodiments except for anthropomorphic form can be considered. Compared with anthropomorphic robots, the other three categories of robots have another advantage. That is their social expectation is lower than that of anthropomorphic

robots such that their interaction skills with humans are easier and simpler.

2.1.2 Representative Social Robotics for A Child

In Chapter 1, we have reviewed three representative social robotics for a child. They are AIBO for entertainment, Probo for healthcare, and PaPeRo for childcare. In addition, there are more other social robotics used in these areas or related areas for a child. Figure 2.2 shows several social robotics for a child. Among these robots, some of them have been commercially available such as AIBO, PaPeRo, QRIO SDR-4X, iRobiQ, Paro, Keepon, and iCat, and others such as Probo, RUBI, Huggable, Engkey, and Iromec are still being developed to assist our daily lives. Generally, these these robots can serve as many functions and the application for a child is one example. Since these robots have demonstrated good performance in children-related areas, we will review them in this chapter.

In the entertainment area, QRIO SDR-4X is another representative robot besides AIBO. It is a small biped robot [39] which is developed by Sony Corporation. It has 38 DOFs, standing 58cm, and can fulfill motion and communication entertainment. There are two main entertainment abilities in SDR-4X, which are dancing and singing. When singing a song, the robot can demonstrate different emotional expressions. In addition, SDR-4X can accomplish several human-like behaviors, such as walking on various floor conditions, human identification, and speech communication by using its visual, audio, and tactile systems [12]. Besides in the home environment for entertainment, the robot has been utilized in an early childhood education center to study socialization between toddlers and robots due to its impressive mechanical and computational skills [40].

Figure 2.2: Several representative social robotics for a child. From left to right and top to down, they are AIBO [11], Probo [13], PaPeRo [15], SDR [11], RUBI [42], iRobiQ [44], Paro [45], Huggable [24], Keepon [47], iCat [48], EngKey [49], and Iromec [50], respectively.

In the education area, one typical example is RUBI. RUBI is a three-feet tall robot. It consists of a head, two arms and a touch screen, and is designed to assist teachers for early childhood education. RUBI was set at the Early Childhood Education Center at the University of California, San Diego, to interact with the children with 18-24 months old. It can teach children numbers, colors and some basic concepts, and schedule proper lessons and assist teachers according to the children's emotional responses [41]. RUBI contains some perception functions

such as face detection and tracking, and emotion recognition [42, 43].

Similar to RUBI, iRobiQ is another robot designed for children's education by Yujin Robot Co.,Ltd., and has been commercialized recently [44]. The robot is about 45cm in height and its head, arms, and wheels can move. In addition, it can express simulated emotions by using face lamps. There are mainly four menus used in iRobiQ, which are Thematic learning, English, Playground, and Teacher's room. Children can select and use these menus by touching them on the screen which is mounted on the robot's belly. Through the designed functions in these menus, children can learn contents by theme, study English, listen stories, and play puzzles together. While for the teachers, the robot is able to help them to check attendance of children and play study materials.

In addition to entertainment and education, social robotics have been employed for children therapy, such as therapy for a child with Autism. Paro is such a representative robot [45]. This robot is designed with an appearance of a baby harp seal that is covered with pure white fur. When humans hug the robot, the contact with Paro can be measured by ubiquitous surface tactile sensors of the robot. By analyzing the collected signals, the robot gives proper response according to different touching from humans. Besides tactile sensors, Paro also uses a visual sensor to sense light, an audio sensor to localize sound source and recognize speech, and a balance sensor to adjust its movements. To extend interaction time with the robot, Paro has the ability to demonstrate the preferred behaviors of its owner when it lives with its owner in a long time. Due to its physical interaction with tangibility, Paro has been applied to the therapy of children [46].

Huggable [25] is another representative robot used for children therapy. Similar to Paro, it mainly utilizes tactile-based signals to sense the outside environment. Huggable has the appearance of Teddy bear, and is covered with a full-body sensitive skin containing more than 1500 sensors. Hence, it can detect and recognize pressure from the outside world. In addition, cameras and microphones are used. After semantically analyzing the collected data, the robot can convey a personality-rich character through some gestures and expressions. Moreover, it can be remotely controlled and applied to monitor the elders and children through a web interface. Due to these impressive features, Huggable is also applied for healthcare, education, and social communication.

Besides the above mentioned robots, there are several other social robots which can be applied in children-related areas. For example, Keepon [47], a small yellow snowman with a black cylinder, was developed to study social development in research institutes, assist autism therapy in care centers, and play with children in a playroom; iCat [48], a cartoon cat without mobile ability, was designed to be a family companion to control homely used devices and play games with a child; Engkey [49], a spheroid robot with head, arms, and wheels, was developed to provide educational assistance to native and Korean teachers in teaching the English language to students; and Iromec [50], a modular robot including a mobile platform, an interaction module, and some control buttons, was designed to engage in social exchanges with different disable children like Autistic children, Moderate Mentally Retarded children, and Server Motor Impaired children.

2.1.3 Discussion

By observing the embodied forms of the reviewed robots in Figure 2.2, it can be seen that small humanoid robot SDR-4X belongs to anthropomorphic robots; AIBO and Paro are examples of zoomorphic robots; Probo, PaPeRo, RUBI, iRobiQ, Huggable, Keepon, iCat, and Engkey are caricatured robots; and Iromec is a functional robot. Most of these robots belong to caricatured robots. It implies that when designing a social robot for children, caricatured representation may be a good choice because it adheres to Mori's "uncanny valley" hypothesis [18]. Specifically, due to less human likeness of these robots, they can avoid uncanny valley shown in Figure 2.1. While these robots cannot reach the first peak in the figure which is in 100% human likeness, they can reach the second peak which is in 70% human likeness by a suitable design, where the peak value refers to acceptance degree to the robot among humans. Moreover, due to unusual embodied forms of caricatured robots, several unrealistic functions could be designed for desirable tasks. For instance, if a touch screen is mounted on the robot's belly, some functions like video playing can be included. Additionally, the child can easily operate the robot by touching the screen. Probo, RUBI, and iRobiQ are three representative examples.

As we mentioned above, different design objectives have different design methods. Consequentially, the developed robots will demonstrate different appearances, functions, and interaction ways. Even for similar applications, there will be large differences in the built robots. That is because distinct designers may have distinct understanding to their developed robot systems and different design ideas

will lead to distinct forms of robots. Therefore, it is difficult to simultaneously apply the same robot in different areas. To effectively satisfy the design objectives, robotic researchers can develop their own robots by following proper design methods and taking existing robots as references. Besides building the robots by self, there is another situation. For some researchers who just use robots to study or test some theories or algorithms, they can directly utilize or slightly modify the existing robots. In this study, we aim to develop a robotic nanny to play with and take care of a child during his/her parent or caregiver absences. While the above reviewed robots show good performance in their target environments, they cannot be directly applied as a robotic nanny, or cannot satisfy our design objectives. They can only be used as references. Hence, we develop our own robot. Having reviewed design approaches and several representative social robots, we choose functionally designed methods and caricatured form for our robot.

2.2 Facial Expression-Based Emotion Recognition

Emotion recognition plays an important role in social robotics for a child. To recognize a user's emotions, there are several cues that can be utilized, such as speech, facial expressions, and gestures. Since Mehrabian [26] has shown that in human face-to-face communications, 7% and 38% information are transferred by spoken language and paralanguage, and 55% is transferred by facial expressions, we select facial expressions to recognize emotions in this study. Besides emotion recognition, facial expressions have been used to study social interaction, mental activities, and physiological signals. Due to the wide applications, a large number

of research work related to facial expression analysis has been concluded.

For facial expression-based emotion recognition, there is a long history going back into the nineteenth century. A pioneering work was Darwin in 1872 [51] who referred to the universality and continuity of facial expressions in man and animals, and stated the relationships between some inborn emotions and serviceable associated habits. Motivated by Darwin's work, Ekman and Friesen in 1971 [52] proposed that six basic emotion states including happiness, sadness, fear, disgust, surprise, and anger can be expressed by a unique and universal facial expression under different human ethnicities and cultures. In 1978, Ekman and Friesen [53] produced a facial action coding system (FACS) to recognize emotion by different facial expressions. FACS described "all visually distinguishable facial movements" caused by action units (AUs). There are 46 AUs used in the coding system to express different facial movements. Based on FACS, Ekman et al. developed a technique called Emotion FACS (EMFACS) [54] to score certain key AUs which are relevant to detect emotion. It provided an effective way to reduce scoring time for the researchers who focus on facial emotion signals. To conveniently link facial expressions with their psychological interpretations, Ekman et al. built a Facial Action Coding System Affect Interpretation Dictionary (FACSAID) [55]. It can well describe the relationships of FACS scores, facial behaviors, and expressed emotions.

Due to the efficacy and convenience of FACS to describe different facial movements and emotions, this system has been widely used in facial expression analysis and synthesis, and becomes a baseline of extensive facial expression recognition methods. These methods generally require to locate characteristic facial regions like

forehead, eyes, cheeks, nose, and mouth, and extract facial features from these regions such as meaningful points and lines which represent the movements and shapes of eyes, nose, and mouth.

Inspired by the importance and efficacy of facial expressions in emotion, lots of researchers have shown great interest in the problem of detecting emotion from facial expressions. Over the past three decades, a large number of related methods have been proposed [23, 56, 57, 58], and they can be divided into two main categories: geometry-based and appearance-based. Geometry-based methods usually extract facial features such as the shapes and locations of facial components (like the mouth, eyes, brows and nose) and represent them by a feature vector to characterize the facial geometry [27, 28]. The above FACS-based methods belong to this category. While the geometry-based methods can well interpret facial expressions and emotions and have shown reasonable performance under controlled environments, it is very challenging to precisely localize and extract these features in many practical applications such as human-robot interaction due to complex backgrounds and varying illuminations. Hence, we choose appearance-based methods to recognize facial expressions. Appearance-based methods are popular for facial expression recognition and also demonstrate reasonable performance in terms of the recognition accuracy.

2.2.1 Appearance-Based Facial Expression Recognition

Appearance-based methods holistically convert each facial image or specific facial regions into a feature vector and then apply image filters or some learning

Figure 2.3: Emotion-specified facial expressions which are anger, disgust, fear, happy, sad, surprise, and neutral expressions, respectively [56].

techniques to extract some discriminative features for facial expression representation [29, 30]. These methods generally extract features such as local binary pattern (LBP) feature, intensity feature, Haar-like feature, and Gabor wavelet feature. Based on these extracted features, the tested specific facial regions are classified into corresponding facial action units and the whole tested facial images are labeled with prototypic emotional expressions. Some popular classification methods include the nearest neighbor classifier, neural networks, hidden Markov models, and support vector machines. Since Ekman and Friesen [52] claimed that prototypic emotional expressions are universal under different human ethnicities and cultures, most facial expression methods attempt to recognize these basic emotional expressions that are comprised of anger, disgust, fear, happy, sad, surprise, and neutral expressions, as shown in Figure 2.3, where facial images from the Cohn-Kanade (CK) face database are used [59].

The work of Littlewort *et al.* [60] is an example to recognize 6 basic emotional expressions plus neutral expression. They chose Gabor magnitude to represent facial images. First, the authors convolved the image with a bank of Gabor filters consisting of 8 orientations and 5 spatial frequencies. Then they compared the performance of feature selection methods including principle component analysis (PCA) and AdaBoost and recognition algorithms like support vector machine

(SVM) and AdaBoost. Since SVM and AdaBoost only make binary decisions, better decision strategies for multiclass classification should be used. Here, the authors evaluated K-nearest neighborhood, a voting scheme, and multinomial logistic ridge regression (MLR). Experiments conducted on the CK and Pictures of Facial Affect (POFA) databases have shown that the combination of AdaBoost as a feature selection method, SVM as a classification algorithm, and voting as a multiclass decision strategy can obtain better recognition accuracy. Furthermore, the presented recognition system can also be used to recognize facial action units.

In addition to Gabor wavelet feature, local binary pattern (LBP) is another appearance feature which is originally presented and applied to texture analysis. Due to its strong tolerance to lighting changes and computational simplicity which are very important for real-world applications, it has been widely applied for facial expression analysis. Shan and colleagues [61] performed person-independent facial expression recognition by utilizing LBP features. Template matching, SVM, linear discriminant analysis (LDA) and linear programming techniques were chosen as the classification algorithms. Since LBP feature is a histogram to statically describe the characteristics of an image, Chi square distance was deployed in template matching. Experimental results on the CK database have shown that SVM obtains the best results. The authors also proposed boosting LBP that was learned by Adaboost and can further improve the recognition performance as it contains more discriminative information to represent facial images. Then SVM was deployed to recognize facial expressions. The results have shown that it can achieve better recognition accuracy than that obtained by using only SVM. Moreover, the authors evaluated its generalization ability on another two databases:

MMI database and JAFFE database. The accuracy rates about 6 basic expressions and neutral expression are only 51.1% for MMI and 41.3% for JAFFE by using boosted-LBP and SVM (RBF).

Based on the LBP feature, Zhao and Pietik*ä*inen [62] presented a spatiotemporal LBP (LBP-TOP) method which extends the original LBP on three orthogonal planes including XY, XT and YT for facial expression recognition, in which X and Y are the width and the height of each face image, and T is the length of image sequences. The proposed feature not only has the original feature's advantages like the robustness to illumination variation, but also can represent facial expression's temporal characteristics. The proposed video-based LBP feature with AdaBoost as a feature selection algorithm and SVM as a classification method obtained good accuracy results on the CK database. Moreover, it can be used in real-world environments.

Actually, the proposed spatiotemporal LBP is different from other features described above. This is because it is a video-based feature, and others are image-based. As we know, when humans show their facial expressions, facial expression may change over time. Thus the temporal information resulted from the change could well describe dynamic facial features and is significant to distinguish various facial expressions. More and more researchers have realized it and put more attention on video-based features. Yang and colleagues [63] proposed an encoded dynamic feature to represent facial images. Due to lower computation cost of Haar-like features, they were selected to be dynamic features by following two steps: first, the whole image is described by Haar-like features, and then features from consecutive frames are combined. Inspired by LBP features, the dynamic

Haar-like features were encoded into the binary patterns in terms of a code book. Finally, AdaBoost was applied to recognize facial expression. The experimental results based on the CK database for 6 basic facial expressions have shown that the proposed features can achieve better results compared with Gabor wavelet feature in the form of the area under the receiver operating characteristic (ROC) curves. Moreover, it can obtain a promising performance when used for action units recognition.

Besides recognizing prototypic emotional expressions, appearance-based methods are employed to detect facial action units. Donato *et al.* [64] applied optic flow, PCA, local feature analysis, LDA, independent component analysis (ICA), local PCA, and Gabor wavelet filter to recognize action units in upper and lower faces, where the nearest neighborhood and template matching classifiers were used. Experimental results were compared with those of humans and the results have shown that Gabor wavelet representation can obtain the best result.

To clearly demonstrate the experimental settings and performance of each method introduced in this subsection, we tabulate the extracted features, classification methods, recognition accuracies, emotion categories, training and testing settings, and the employed databases of these methods in Table 2.1. It can be seen from the table that each method has shown good performance in terms of recognition accuracy under their experimental settings. However, these enumerated methods cannot be directly compared according to the recognition accuracy listed in the papers. The reason is that these methods were conducted on different databases. In the absence of comparative tests on common data, it is difficult to determine the relative advantages and disadvantages of different approaches.

Table 2.1: The methods for facial expression analysis described in this subsection.

Reference	Feature	Method	Accuracy	Categories	Train cases	Test cases	Database
Littlewort [60]	Gabor wavelet filter	AdaBoost + SVM + voting	93.3%	7 class	leave-one-out cross-validation (626 from 90 subjects)		Cohn-Kanade
			97.3%	7 class	leave-one-out cross-validation (110 from 14 subjects)		POFA
Shan [61]	Boosted-LBP features	SVM	1)91.4% 2)95.1%	1)7 class 2)6 basic emotions	10-fold cross-validation (1280 from 96 subjects)		Cohn-Kanade
			86.9%	7 class	10-fold cross-validation (384 from 20 subjects)		MMI
			81.0%	7 class	10-fold cross-validation (213)		JAFFE
Zhao [62]	LBP-TOP	Adaboost + SVM	93.85%	6 basic emotions	2-fold (374 sequences from 97 subjects)		Cohn-Kanade
Yang [63]	encoded dynamic Haar-like features	AdaBoost	A:0.982, D:0.987, F:0.83, H:0.983, Sa:0.946, Su:0.996 (ROC)	6 basic emotions	60 subjects	36 subjects	Cohn-Kanade
Donato [64]	ICA or Gabor wavelet filter	Nearest neighborhood	95.5%	AU 1, 2, 4, 5, 6, 7 (upper face) AU 17, 18, 9+25, 10+25, 16+25, 20+25 (lower face)	leave-one-out cross-validation (111 sequences from 20 subjects)		Ekman-Hager

Table 2.2: Generalization performance to independent databases.

Reference	Accuracy	Train database	Test database
Littlewort [60]	60.0%	Cohn-Kanade	POFA
Shan [61]	51.1%	Cohn-Kanade	MMI
	41.3%	Cohn-Kanade	JAFFE

Among these methods, Littlewort *et al.* [60] and Shan *et al.* [61] also tested the approaches' generality by using two different databases as training data and testing data, respectively. Table 2.2 lists the results of two methods. From the table, we find that the recognition accuracies on two databases drop heavily when compared with those obtained on the same database. This is because training and testing data from the same database are usually collected under the same condition such that they are independent and identically distributed. It is easier to obtain good performance for the proposed approaches under such condition. While for the training and testing data from different databases, there is big variance between them such that the performance of the proposed approaches will be affected. This is also called "cross-dataset" recognition problem which is universal in real-world facial expression recognition. It is a challenging problem, and has been deemphasized in this area.

2.2.2 Facial Expression Recognition in Social Robotics

Facial expression recognition has been employed in several real-world applications to recognize humans' emotional states. Human-robot interaction in social robots is a representative application. For example, among the reviewed social robots in section 2.1, RUBI has the function of facial expression recognition. The developed system can first automatically detect frontal faces in the video stream and then code each facial image with 20 action units. Based on the detected facial images, it firstly extracted Gabor wavelet features, and then chose Adaboost as feature selection method and SVM as data-driven classifier. In addition to the posed expression databases, the developed system also showed good performance

Table 2.3: Properties of an ideal automatic facial expression recognition system.

Robustness	
Rb1	Deal with subjects of different age, gender, ethnicity
Rb2	Handle lighting changes
Rb3	Handle large head motion
Rb4	Handle occlusion
Rb5	Handle different image resolution
Rb6	Recognize all possible expressions
Rb7	Recognize expressions with different intensity
Rb8	Recognize asymmetrical expressions
Rb9	Recognize spontaneous expressions
Automatic process	
Am1	Automatic face acquisition
Am2	Automatic facial feature extraction
Am3	Automatic expression recognition
Real-time process	
Rt1	Real-time face acquisition
Rt2	Real-time facial feature extraction
Rt3	Real-time expression recognition
Autonomic process	
An1	Output recognition with confidence
An2	Adaptive to different level outputs based on input images

in spontaneous expressions [43].

Similar to normal commercialized products, the final goal of social robots is that the developed robots should be able to perform with less human interference. In addition, they should be capable of providing correct and real-time responses to their users. To satisfy these requirements, the used facial expression recognition system must perform automatically and in real time. Moreover, it should be able to output recognition results with high confidence under various and complex environments. Table 2.3 lists the properties of an ideal facial expression analysis system proposed by Tian et al. [65].

By observing and analyzing each property in the table, we can find what social

robots require is an ideal facial expression analysis system. However, each property is a challenging problem in facial expression recognition area and it still requires the researchers from different disciplines to address these challenges. Compared with the social robots used in outside environments, these challenging problems are relatively easy to be solved for a robotic nanny because the robotic nanny is usually applied in home, and the application environment may be simpler. Even so, more efforts should be made to drive the current facial expression recognition techniques towards a practical robotic nanny. Note that we in this study only discuss feature extraction and expression recognition. For an automatic facial expression recognition system in social robots, it should also include face acquisition, which detects facial images from input images and is a preprocessing stage before feature extraction step. We will not discuss this problem in detail.

2.2.3 Discussion

By reviewing several representative facial expression recognition methods and analyzing the properties of the above mentioned facial expression analysis system, we conclude that even if many existing methods have achieved satisfactory recognition results under their experimental settings, there is still some room to improve them for real-world applications. This is because their experiments were conducted under controlled conditions and did not consider some real-world factors such as individual difference in subjects, distinct data collection scene, out-of-plane head motion, and more spontaneous expressions. If these existing facial expression recognition methods are directly used in real-world applications without any improvement, their performance will undoubtedly drop.

In Section 2.2.1, some researchers have studied cross-dataset problem in facial expression recognition. The unsatisfactory results showed that the existing methods are difficult to achieve promising performance on this problem. Cross-dataset problem refers to that the training and testing data used in experiments are collected under different conditions. Generally, the testing data are collected online and may be different from the training data, such as different races, illuminations and imaging conditions. This is popular in real-world applications. However, to the best of our knowledge, with some exception work such as face recognition [66] and age estimation [67], this problem is seldom addressed in the literature on facial expression recognition.

Another problem that is significant in automatic facial expression recognition system and seldom studied by researchers is how to develop a misalignment-robust facial expression recognition method. As we know, most existing appearance-based facial expression recognition methods can only work well when face images are well-aligned. However, under uncontrolled conditions, it is very challenging to obtain well-aligned face images due to the eye localization error in automatic face acquisition and alignment procedures. The caused spatial misalignments will unavoidably affect the performance of these appearance-based facial expression recognition methods. This has been proved by Gritti et al. by investigating several local features based facial expression recognition methods [68].

While for other real-world factors such as out-of-plane head motion and spontaneous facial expression recognition, since they directly influence the real-world applications of facial expression recognition system, more and more researchers have been made to handle them. For example, to deal with out-of-plane head

motion, action appearance models [69], local parametric models [70], 3D motion models [71], and feature point tracking techniques [72] have been proposed; to analyze spontaneous facial expression, the systems developed by CMU [73] and UCSD [74] have achieved some promising results through recognizing a few action units.

In this thesis, we will mainly focus on studying cross-dataset facial expression recognition and misalignment-robust facial expression recognition from theoretical aspect. This is because facial images captured across datasets and with misalignments usually occur in facial expression recognition in real-world applications. Investigating these two problems can effectively improve the performance of an automatic facial expression recognition system. Current publicly available databases for facial expression recognition are only for adults and not for children, and Sullivan and Lewis [75] have shown that children are similar to adults when they display different facial expressions. Hence, we use facial expression databases developed for adults to study the above mentioned two problems.

Chapter 3

Design and Development of A Robotic Nanny

3.1 Introduction

With the rapid development of current society, parents become more busy and cannot always stay with their children. Hence, a robotic nanny which can care for and play with the children is desirable. A robotic nanny is a subclass of social robots acting as a child's caregiver [8] and aims to extend the length of parents or caregiver absences by providing entertainment to the child, tutoring the child, keeping the child from physical harm, and ideally, building a companionship with the child [9, 6]. With the help of a robotic nanny, it can release the pressure on parents to a certain extent. Furthermore, due to the concentration of high technologies in the robot, it may activate the child's interest to play with the robot and learn new knowledge during the interaction.

Currently, a large number of social robotics have been developed for children in entertainment, healthcare, and domestic areas [2]. For example, Sony's dog

robot AIBO is designed to be a robotic companion/pet of a child [12]. RUBI is designed to be an assistant of teacher for early childhood education [42]. Probo is developed to emotionally interact with the children in a hospital [14]. NEC's PaPeRo is developed to care for children in domestic and public environments [16]. Some of these robots are commercially available such as AIBO and PaPeRo, and others are being developed. While these robots have demonstrated promising performance in their target environments, they cannot be directly applied as a robotic nanny like AIBO, or cannot satisfy our specific design objectives such as Probo, RUBI, and PaPeRo.

In this chapter, we develop a robotic nanny named Dorothy Robotubby to play with and take care of a child during his/her parent or caregiver absence. We expect our robot not only interacts with a child, but also builds a connection between a child and his/her parent. When interacting with the robot, it is hopeful that the child will not get bored in a short term. The developed robotic nanny is specified for a normal child and used at home.

To achieve this goal, there are two main problems to be addressed in our robot. The first is how to activate and maintain a child's interest and curiosity to interact with the robot. We solve this problem by designing a robot with a caricatured appearance, various functions, and a conveniently-operable user interface. Among these three design issues, the developed functions include storytelling, playing music, games, and chatting. When executing these functions, the robot can show different facial expressions and gestures. In addition to these direct interaction ways with the robot, two functions including face tracking and emotion recognition continuously work during the whole procedure of the robot operation. Compared

with TV or several interactive computer softwares which can maintain a child's interest and curiosity, the robot can not only achieve similar objectives such as playing videos or games, but also implement several functions related to actions or motions such as face tracking and physical interaction. This will keep a child entertained, learning, safe, and makes the robot more human-like and acceptable.

The second problem is how to connect a child and his/her parent via the robot. According to several literature on psychology [6], while direct evidence for the harm of parent/caregiver absences to a child is still lacking, it is undoubt that parents/caregivers play an important role in a child's development. Hence, a robotic nanny cannot totally substitute parents, and it would be better to provide a natural way to keep the child and his/her parents in touch. A feasible solution is to transmit a child's and parent's video and audio data to each other by video calling. Different from conventional video calling functions, when a child is talking with his/her parent through the robot, the parent can remotely control the robot to execute several commands such as showing different facial expressions and giving a remote hug. This can help parents to express their emotions to their child in a physical way through the robot. Moreover, the robot can continuously transfer what it sees to parents through images such that it can keep a child from physical harm under a parent's surveillance.

This chapter is organized as follows. Section 3.2 overviews our robotic nanny – Dorothy Robotubby. Section 3.3 introduces two main user interfaces in the developed robot system. Section 3.4 describes each function of Robotubby, and Section 3.5 concludes this chapter.

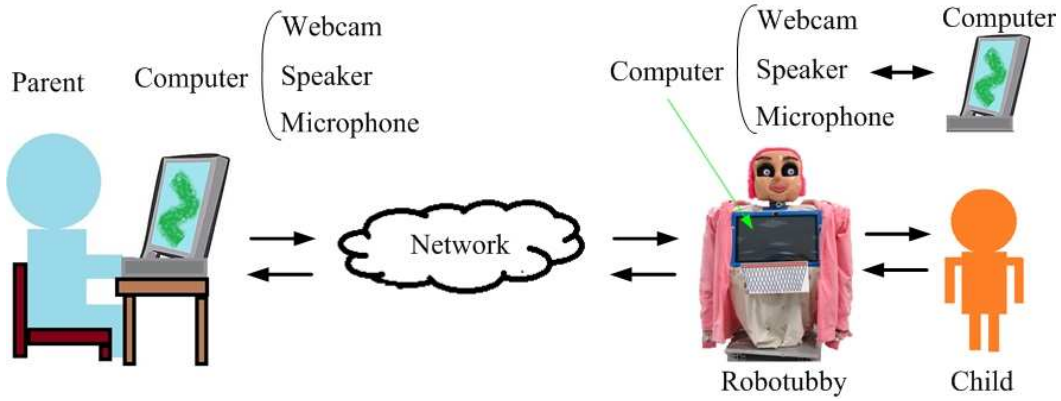


Figure 3.1: System configuration

3.2 Overview of Dorothy Robotubby System

3.2.1 System Configuration

To develop an acceptable robotic nanny, two aspects are highlighted in our robot – Dorothy Robotubby. One focuses on the design of robot itself, mainly including appearance, function, and user interface designs. The other focuses on remote control of parents. It mainly consists of the function and user interface designs. To guarantee such a robotic nanny system to work well, several engineering technologies should be integrated, such as a robot's morphology, mechanical and electrical designs, and software development.

Figure 3.1 shows the system configuration of Dorothy Robotubby. We can see that there are three computers used in the whole system. One touch screen computer is mounted on the belly of the robot and utilized to control the robot. The child can interact with the robot by directly clicking buttons on the touch screen. The second computer is employed to help the robot to accomplish some complex tasks such as emotion recognition. After computation, the results are sent back to

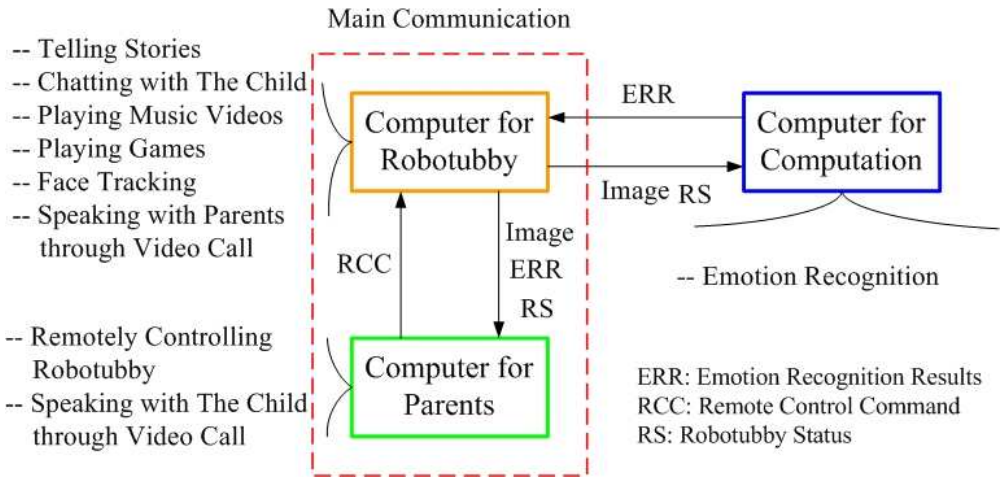


Figure 3.2: Schematics of the whole system

the robot via the local network. These complex computation tasks require more computer resources and thus may affect the accomplishment of other functions of the robot. We solve this problem by using another computer. The second computer is typically located at home, connected by a local network within the home. The third computer is used by parents to remotely control the robot. It can transmit information to the robot through the network. Besides three computers and the robot, two webcams, one microphone, and one speaker are installed in the system for video calling and surveillance.

Figure 3.2 illustrates the corresponding schematics of the whole system. From the figure, we can see that the computer in Robotubby is the server, and the other two computers are the clients. The computer in Robotubby transfers images of a child to the computer of parents for surveillance and to the other computer for computation, respectively. It also transfers the robot's status to the other two computers to assist network connection and remote control. Conversely, the computer for computation transfers emotion recognition results to the computer

of the robot, and the computer of parents transfers remote control commands to the computer of the robot.

Besides transferring information among three computers, each computer has its own tasks. Specifically, by controlling the computer of Robotubby, the robot can tell stories, chat, play music videos and games, and track a child's face. Additionally, a child can speak with his/her parent though video calling of the robot. The main task of the computer for computation is to help the robot to recognize the child's emotions. While for the computer of parents, in addition to remotely controlling Robotubby, the parent can talk with his/her child via video calling function.

3.2.2 Dorothy Robotubby Introduction

Dorothy Robotubby is developed to be a robotic nanny for a child. We design Robotubby with a upper body and a caricatured appearance. It mainly consists of a head, a neck, a body, two arms, two hands, and a touch screen in its belly. Figure 3.3 shows the main components of Robotubby.

Robotubby uses different input and output (I/O) devices to connect with the outside environment. Table 3.1 lists the used devices in our robot. Specifically, two web cameras provide images for face tracking, emotion recognition, and video call functions. One microphone collects audio signals for the video call function. Two limit switches acquire tactile signals for game playing function. These devices are the main input devices. For the output devices, there are one speaker, 14 Dynamixel servos, and 9 Lynxmotion servos. The speaker is used to produce

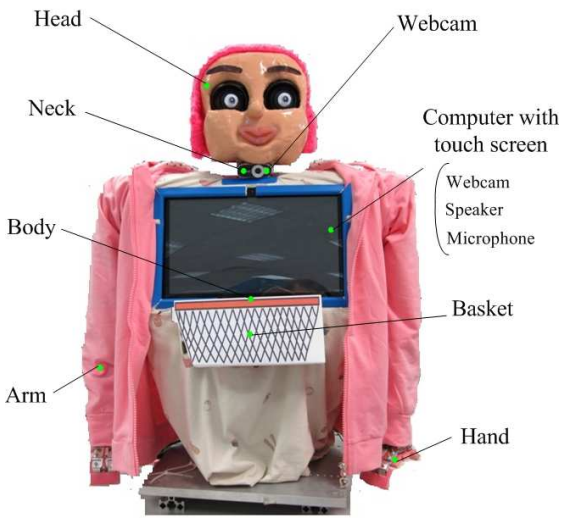


Figure 3.3: Main components of Dorothy Robotubby

Table 3.1: Input and Output Devices.

Devices	Quantity	Input / Output
Web Camera	2	Input
Microphone	1	Input
Limit Switch	2	Input
Speaker	1	Output
Dynamixel Servo	14	Output
Lynxmotion Servo	9	Output
Touch Screen	1	Input / Output

sounds in Robotubby. Dynamixel servos and Lynxmotion servos are employed to control the body and head of the robot, respectively. In addition, another important device is touch screen which can be used for both input and output. When the child operates the robot by clicking buttons on the touch screen, the touch screen is an input device. While when the robot gives some responses such as playing music videos through the touch screen, the touch screen is an output device.

Table 3.2: The information of a Samsung Slate PC.

Type	Samsung Slate 7
Processor	Intel Core i5 2467M
RAM	4GB DDR3 1333MHz
Hard Disk	128GB SSD HDD
Operating System	Windows 7 Professional
Screen	11.6 PLS LCD 16.7M Colour Capacitive Touch screen
Resolution	1366 X 768
Dimension	296X184X13 mm
Weight	900g

The touch screen in Robotubby is a part of a Samsung Slate PC. The reason to select this computer in our robot is a comprehensive consideration of the size and weight of the robot, the execution speed of the robot's tasks, and the development convenience of touch screen technology. The detailed information on this PC is shown in Table 3.2. Due to the fast speed of the processor and large number of RAM, this Slate PC is also a controller to connect and control all the I/O devices listed in Table 3.1 to execute the tasks of Robotubby. These tasks are described by predefined codes which are programmed using C♯ language.

To activate a child's interest to interact with the robot, Robotubby is designed to be able to demonstrate different facial expressions and gestures by controlling its face and body components when it executes several tasks. Specifically, there are 9 Degrees of Freedom (DOF) in the robot's head, 2 DOF in its neck, 2 DOF in its body, 2 DOF in its each shoulder, 2 DOF in its each arm, and 1 DOF in its each hand. The correspondingly employed servo motors are listed in Table 3.3. Among these servo motors, HS-65HB uses SSC-32 servo controller from LynxMotion which communicates with the computer using RS-232 signals. The rest motors directly utilize the computer as their controller and connect with the

Table 3.3: The used servo motors in Robotubby.

	Type	Quantity
Head	HS-65HB	9
Neck	RX-24F	2
Body	RX-64	2
Shoulder	RX-64	4
Arm	RX-28	4
Hand	DX-117	2

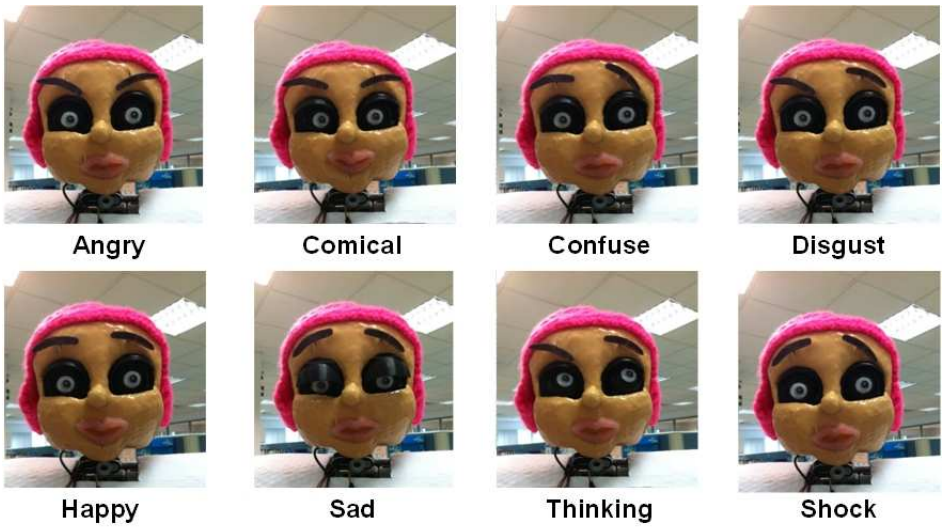


Figure 3.4: Several examples of different facial expressions of Robotubby

computer through RS-485 communication signals. Figure 3.4 illustrates several samples of Robotubby's facial expressions which include angry, comical, confuse, disgust, happy, sad, thinking, and shock, respectively.

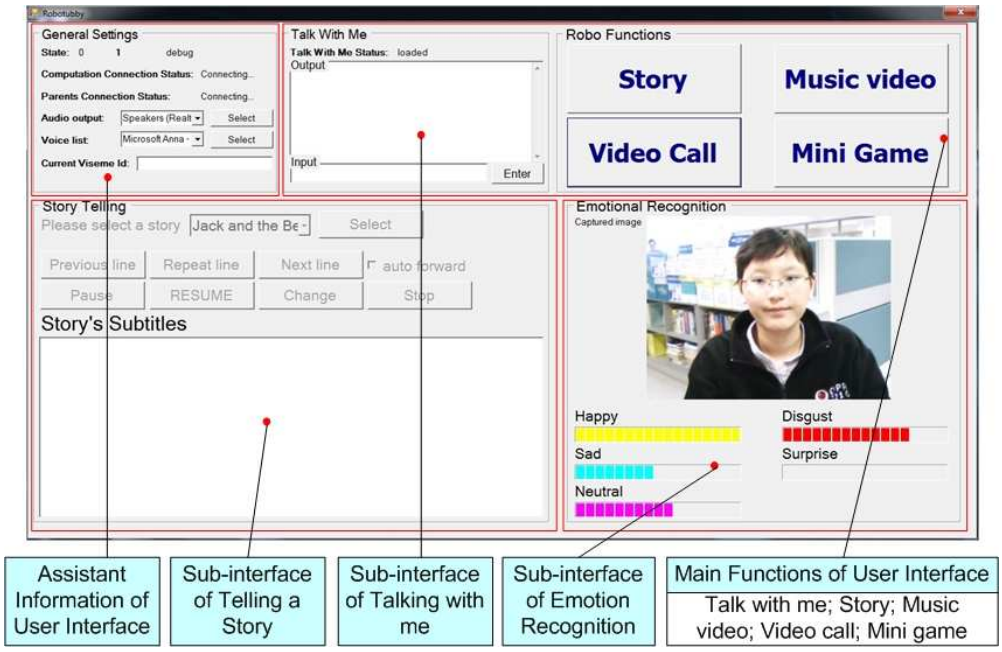


Figure 3.5: User interface of Robotubby

3.3 Dorothy Robotubby User Interface and Remote User Interface

3.3.1 Dorothy Robotubby User Interface

We mount a computer with a touch screen on the robot's belly such that a child can operate the robot by using a mouse or directly clicking buttons on the touch screen. The corresponding user interface is shown in Figure 3.5. From the figure, we can see that the interface mainly consists of five parts: assistant information of user interface, main functions of user interface, sub-interface of telling a story, sub-interface of talking with me, and sub-interface of emotion recognition.

Assistant information of user interface includes working status of Robutubby such as connection status with the other two computers. Normally, it is only used to

provide some information and no further operation from the user is required. The functions of user interface include talking with me, telling a story, playing music videos and games, and video calling. The sub-interfaces of talking with me and telling a story are set in the main user interface shown in Figure 3.5. For the function of talking with me, the child can directly input texts after its sub-interface is loaded. While for the function of telling a story, the child should click "Story" button first. Then he/she can select a story from the list and control it by pressing buttons in its sub-interface. Different from these two functions, other three functions have the independent sub-interfaces. After clicking the buttons of "Music video", "Video Call", and "Mini Game", their own sub-interfaces will pop up.

In addition to the above functions, Robotubby has another two functions: face tracking and emotion recognition. Face tracking begins to work when Robotubby user interface is started. It continuously works until the robot user interface is closed or a stop command is received from the parent's computer. While face tracking function has its own sub-interface, the sub-interface always hides behind the main user interface unless the user forces it to emerge by clicking it. Emotion recognition function runs on the other computer for computation. Once there is a connection between the computer of Robotubby and the computer for computation by network, emotion recognition results of the user will be sent back to the robot and displayed in the sub-interface of emotion recognition in the Robotubby user interface.

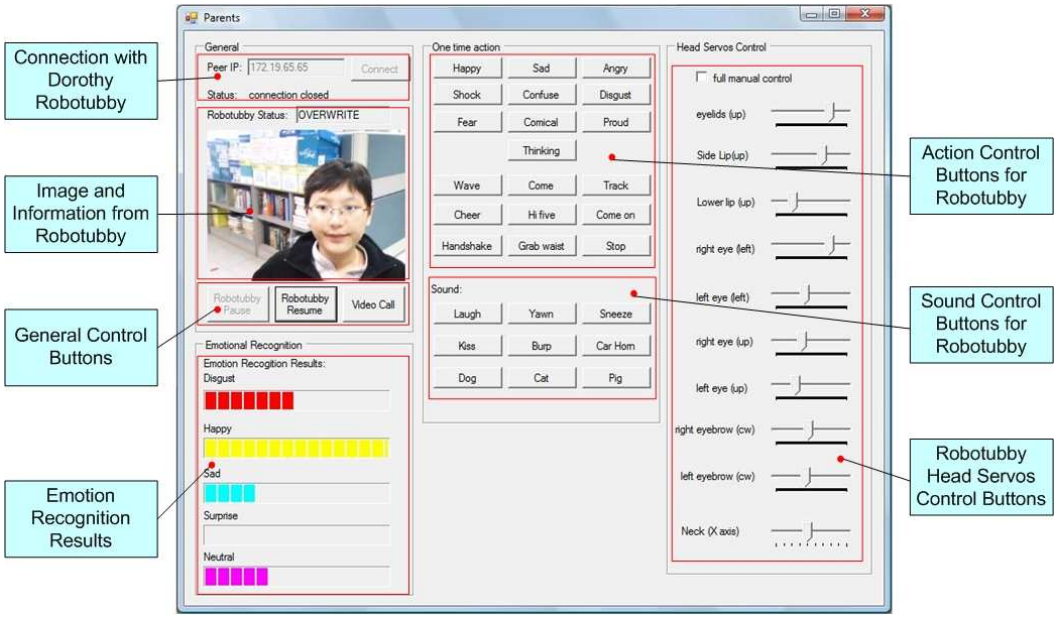


Figure 3.6: Remote user interface

3.3.2 Remote User Interface

In addition to Robotubby user interface, another main interface in our robotic nanny system is the remote user interface which is operated by parents to communicate with their child via network. The main idea to develop this interface is to enhance the connection between a child and his/her parent. Moreover, since the computer of Robotubby can continuously transfer images of the child to the computer of parents, it can keep the child from harm under parents' surveillance. Remote user interface mainly consists of two functions. One is that parents can talk with their child through video calling. The other is that parents can remotely control the robot to execute several commands such as demonstrating different facial expressions and gestures. Figure 3.6 shows the remote user interface with three information categories and four control categories.

Specifically, three types of information include connection with Dorothy Robotubby, images and robot status from Robotubby, and emotion recognition results of a child. Among them, images, robot status, and emotion recognition results are transferred from Robotubby. The transferred images represent what the robot sees. Through them, parents can know what their child is doing, and thus can keep him/her from harm under parents' surveillance. Since there is no camera in current robotic nanny system to capture the robot, through robot status, parents will know which function of Robotubby is operated by the child. While for emotion recognition results of a child, they can provide some reference information to help parents recognize the child's emotional status.

Four control categories consist of general control, action control, sound control, and robot head servo control for Robotubby. General control function is to pause or resume the robot, and call the child through video call. Once "Robotubby Pause" button is clicked, the robot will stop all of its functions except video call and the child cannot operate the robot. Only under such condition, the buttons belonging to other three control categories can be activated and the parent can remotely control the robot through these buttons. If "Robotubby Resume" button is clicked, the parent cannot remotely control the robot except calling the child through video call function and the child can operate the robot again. With respect to action and sound controls, the parent can remotely control the robot to demonstrate several pre-defined facial expressions and gestures such as happiness, anger, and waving hands and play some pre-set sounds such as laugh, yawn, and burp by clicking the corresponding buttons. While for robot head servo control, the parent can remotely and separately control each servo of the robot head by

dragging the sliders representing the different servos.

It should be noted that all these control categories except for sound control can be simultaneously executed with the function of video calling. Hence, when the parent is talking with his/her child through video call, he/she can remotely control the robot to show some facial expressions or gestures. This is different from talking through the telephone or the computer, and it may attract interest of the child.

The main technology problem for developing functions of this interface is how to connect and exchange information with the other computer. Here, we utilized Socket programming in C♯. A Socket is an End-Point of bidirectional (To and From) communication link between two programs including Server program and Client program running on the same network. It has been easily implemented in C♯ through its namespaces like System.Net and System.Net.Sockets.

3.4 Dorothy Robotubby function Description

To attract a child to happily interact with Dorothy Robotubby, we have developed several functions such as telling stories, chatting with a child, playing music videos, playing games, and video calling. These functions of the robot are easily operated by a child. As we introduced in Section 3.3, the child just needs to click the corresponding buttons or input text in the corresponding input textbox in the Robotubby user interface. In addition, there are another two functions including face tracking and emotion recognition which begin to work when Robotubby user interface is started. These two functions work without human intervention. The reason we developed these functions in our robot is that face tracking and emotion

recognition functions can make the robot more natural and believable. Video calling feature allows a more effective communication between parent and child. For the remaining functions, they are useful for the development of children [76, 77] and can be better implemented in our robot. In this section, we will briefly describe each function of the robot.

3.4.1 Face Tracking

When Robotubby user interface is started, face tracking begins to work. It will continuously work unless Robotubby user interface is closed or the robot is remotely controlled by parents. Face tracking aims to track the user's face during interaction with the robot. It utilizes the web camera mounted on the robot neck to capture the image. To accurately detect the user's face in real time, we employed face detector in OpenCV which is an open source for computer vision. The used face detector can real-time detect nearly frontal face with 95% accuracy.

Once the user's face is detected, the position of the face relative to the center of the image can be obtained. The obtained difference will then be used to calculate the required shift for adjusting two servo motors of the robot neck to the correct position. The constant re-aligning will center the user's face and give rise to the face tracking function. Currently, face tracking is only limited to horizontal adjustment. The new positions of two servo motors in the robot neck can be calculated by using Eqs. 3.4.1 and 3.4.2.

$$M1_{NP} = M1_{CP} + D_x/\Delta D \tag{3.4.1}$$

$$M2_{NP} = M2_{CP} + D_x/\Delta D \tag{3.4.2}$$

where $M1_{NP}$ and $M2_{NP}$ refer to the new positions of two servo motors in robot neck, $M1_{CP}$ and $M2_{CP}$ represent the current positions of two servo motors in robot neck, D_x is the difference value between the position of the detected face and the center of the image along x axis, and ΔD is the step width of motor movement relative to the image. To balance the continuity and real-time property of face tracking function, we set ΔD to be 5. If the user's face cannot be detected in several seconds, Robotubby will move the head to its neutral position.

Since the interaction between the user and the robot is mainly face-to-face, the robot can track the user's face in most of time during interaction. Face tracking function can not only make the robot more human-like, but also ensure another function to successfully work. Another function is emotion recognition which relies on frontal facial images to recognize emotion.

3.4.2 Emotion Recognition

Emotion recognition is accomplished by another computer specially used for computation. After two computers for the robot and for computation are connected through network, the recognized emotion results will be shown in Robotubby user interface. Currently, the detected emotional states include happy, sad, disgust, surprise, and neutral expressions. Figure 3.7 shows emotion recognition interface.

Figure 3.7: Emotion recognition interface.

To guarantee the real-time and automatic process of emotion recognition function, we extracted Local Binary Pattern (LBP) feature [61] from eyes and mouth regions of the frontal face. LBP feature uses a histogram of the binary patterns calculated over a region to characterize the texture information of an image. It describes each pixel by the relative gray level information with its neighbors. If the gray-scale value of the neighbor pixel is higher or equal, the value of the described pixel is set to be one, otherwise zero. The descriptor describes the result over the neighborhood as a binary number (binary pattern). Compared with other features like Gabor wavelet, LBP feature is robust to changing illumination and has low computation cost. Moreover, once the used parameters of LBP are determined, it is no need to manually adjust. Hence, it has been widely applied in facial expression recognition.

Motivated by the fact that eyes and mouth regions are most informative when

expressing facial expressions by humans, we applied LBP feature on these two regions rather than the whole face. It ensures that the features extracted from eyes and mouth regions have lower computational cost. To detect these two regions, we employed eyes and mouth detectors in OpenCV. With regard to the recognition classifier, we selected template matching where the template is obtained by averaging the samples in the CK and JAFFE databases and Chi square distance [61] is used to measure the similarity between the template and the testing sample. Due to the simplicity of this classifier, it is easy to guarantee the real-time and automatic process of the emotion recognition system.

Since Dorothy Robotubby aims to be used at home, it means that the user of the robot will be specifically determined. Under such scenario, a feasible solution to improve emotion recognition accuracy is that a classification template exclusively pointing at the user can be trained by capturing five categories of facial expression images of the user before running the robot system. Generally, if the user and environment of the robot system do not change, it is only required to train the template once before using the robot for the first time. If the user do not demand a higher emotion recognition accuracy, the pre-trained template can be used. To provide convenience for training the template by user self, we have developed an independent programme whose interface is shown in Figure 3.8.

In Figure 3.8, the areas labeled with orange boxes are used for collecting facial expression images and training the template. Since the current developed emotion recognition system can detect five categories of facial expressions, the user needs to collect the facial images of each category, respectively. After clicking

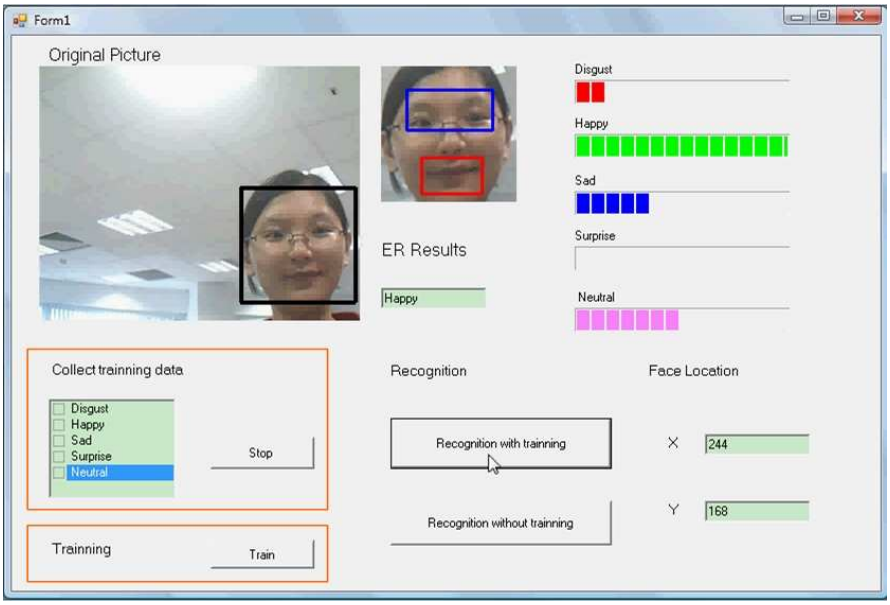


Figure 3.8: Template training interface for emotion recognition.

small square box before each emotion category, the user should display the corresponding facial expression and the program begins to collect and save the images. Clicking the button of "Stop" will stop collecting. After collecting all the images of five categories, clicking the button of "Train" will obtain the template.

3.4.3 Telling Stories

In our robot system, we have prepared five stories that Robotubby can tell. They are "Three pigs", "Little red hood", "Beauty and the Beast", "Jack and the Bean", and "The leap frog". After clicking the "Story" button in Robotubby user interface, the child can select one of them from story list that is shown in Figure 3.9 (a). When Robotubby tells the selected story, the read words will be highlighted by blue color and the buttons of "Previous line", "Repeat line", "Next line", "Pause", "Resume", "Change", and "Stop" can be used to control

Figure 3.9: The sub-interface of storytelling.

Figure 3.10: Several samples of different facial expressions and gestures during telling a story.

the told story. Figure 3.9 (b) shows the sub-interface of storytelling. When telling the story, Robotubby can move its mouth. In addition, we have inserted several specific labels in the prepared stories such that the robot can demonstrate different facial expressions and gestures when meeting them. Several samples of different facial expressions and gestures are shown in Figure 3.10.

Considering that the child may feel boring when listening to a story, the emotional states of the child recognized by emotion recognition function were exploited to determine whether the read story is required to continue. As mentioned above, the

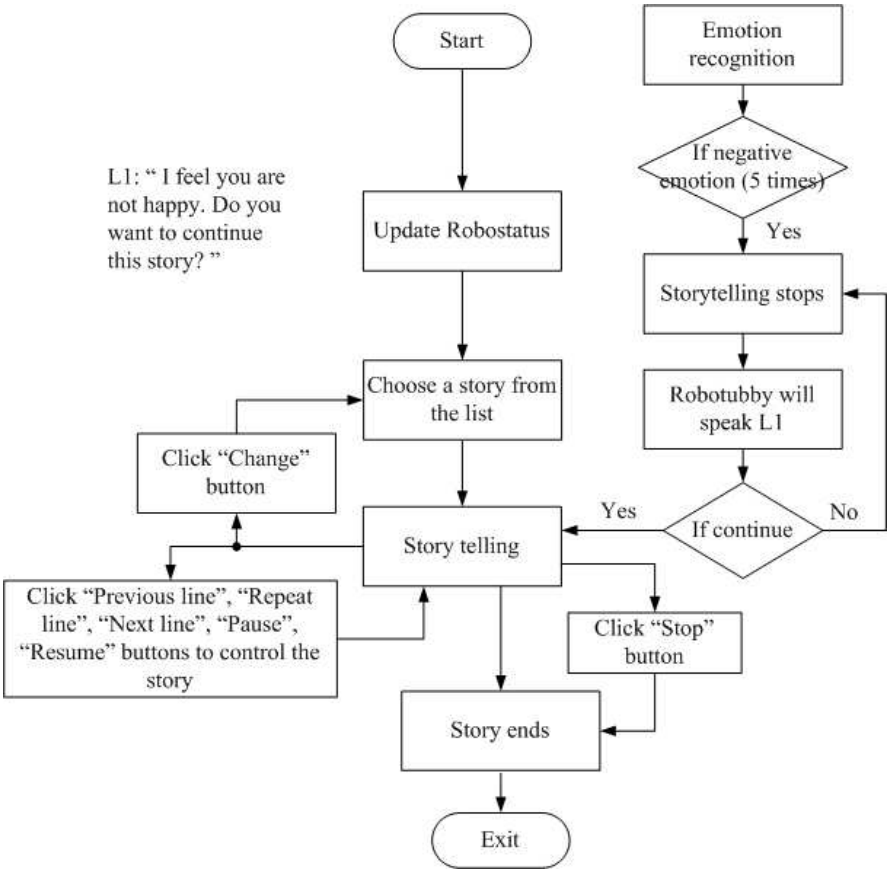


Figure 3.11: The flowchart of storytelling function.

current emotion recognition function can detect happy, sad, disgust, surprise, and neutral expressions. We label sad and disgust expressions as negative emotions. If the detected emotions of the child are negative on five successive times, the robot will pause the on-going story and ask the child whether he/she wants to continue the story. The child can choose to continue the story or change to another function by clicking the buttons on a message box. The whole procedure of storytelling function follows the flowchart shown in Figure 3.11.

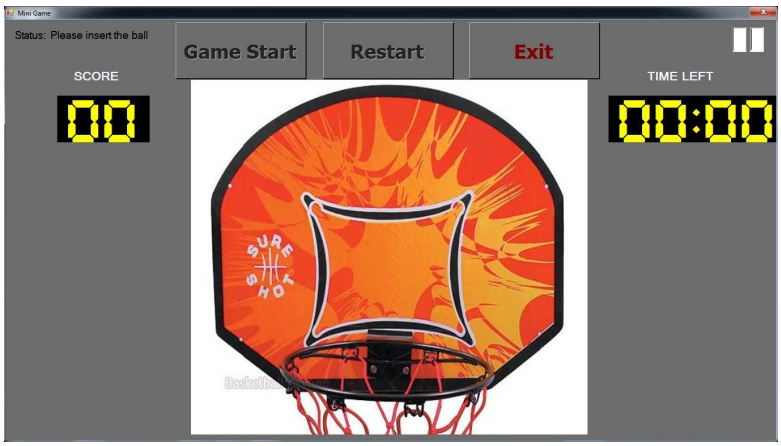


Figure 3.12: The sub-interface of playing games.

Figure 3.13: Several samples of different gestures during the game playing.

3.4.4 Playing Games

This function is to let a child play a simple basketball game with the robot. It is activated by clicking the button of "Mini Game" in Robotubby user interface. The basket mounted on the robot is designed for this game. The main procedure of this game is that the child first delivers the ball into the basket, and then the robot picks up the ball and passes it to the child. Next, it repeats the whole procedure within the time limit. In addition to timing, the robot can count the scores of successfully delivering the ball. The sub-interface of this function is shown in Figure 3.12. The buttons of "Game Start", "Restart", and "Exit" are exploited to control the game playing. Figure 3.13 illustrates the main gestures of the robot during the game playing.

(a) (b) (c)

Figure 3.14: Limit Switch and its locations.

To accomplish this function, two limit switches connected to the LynxMotion controller were used to detect the presence of ball at specific locations. A limit switch [78] is a switch operated by the motion of a machine part or presence of an object which is shown in Figure 3.14(a). In particular, one limit switch is mounted on the basket and the other is installed on the gripper of the robot. These two locations are illustrated in Figure 3.14 (b) and (c), respectively. Once these two limit switches are touched by the ball, it will activate the corresponding programme to control the motion of servo motors such that the robot can fulfill predefined actions to pick up the ball and pass it to the child. The whole procedure of playing game function follows the flowchart shown in Figure 3.15.

3.4.5 Playing Music Videos

Playing music videos is another function of Robotubby. Through this function, the robot can play prepared music videos with the predefined facial expressions and gestures. The demonstrated robot movement is synchronized to the tempo of the song. The current music videos include “If you are happy”, “Three Bears”, “Old McDonald had a Farm”, “Twinkle Twinkle Little Star”, and “Twinkle Twinkle Little Star Sing-A-Long”. The child can select a music video from its list which

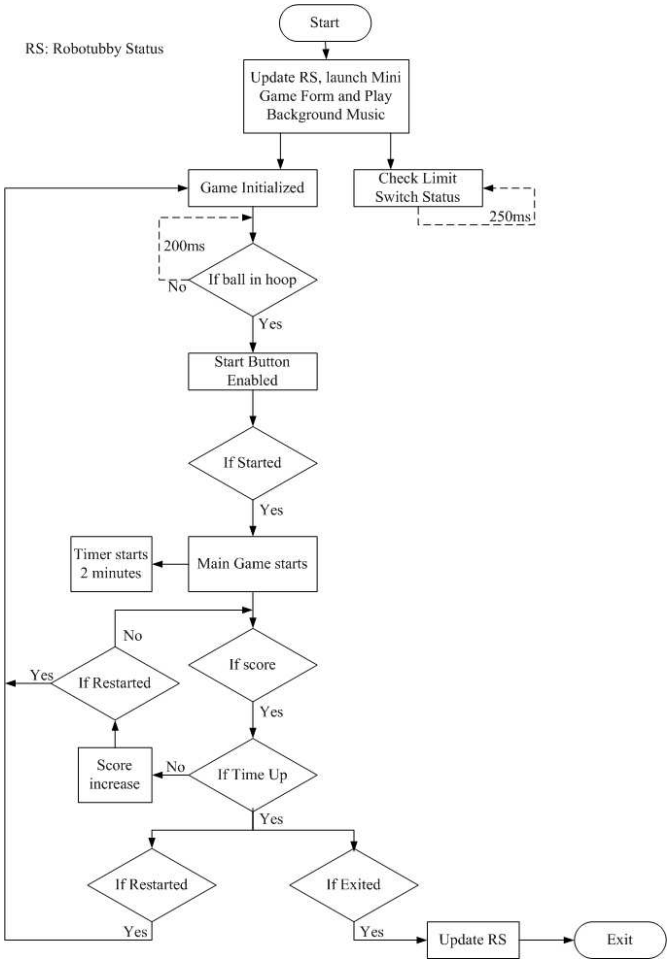


Figure 3.15: The flowchart of playing game function.

is shown in Figure 3.16 (a). Figure 3.16 (b) illustrates the sub-interface of this function. On the interface, the buttons of “Play”, “Stop”, and “Exit” can be utilized to control the music video playing. Figure 3.17 shows several samples of different gestures of the robot during singing a song.

In this function, we incorporated the Windows Media Player console in the sub-interface of playing music videos. Using the functionality of Windows Media Player in a C♯ application can not only guarantee the quality of the played music

Figure 3.16: The sub-interface of playing music videos.

Figure 3.17: Several samples of different gestures during singing a song.

video, but also provide convenience for the development of corresponding programmes. In addition, lyrics of the songs were added such that the child is able to follow and sing along with the music videos and dance along with the robot. This may make the designed function more interesting and entertaining, and enhance the interaction between the child and the robot. The whole procedure of playing music video function follows the flowchart shown in Figure 3.18.

3.4.6 Chatting with A Child

The sub-interface of the function of chatting with a child is illustrated in Figure 3.19. The child can enter his/her question into input textbox. Then the corresponding answer will be given in its output textbox. Figure 3.19 provides two examples of the dialogue between the child and the robot. Similar to storytelling,

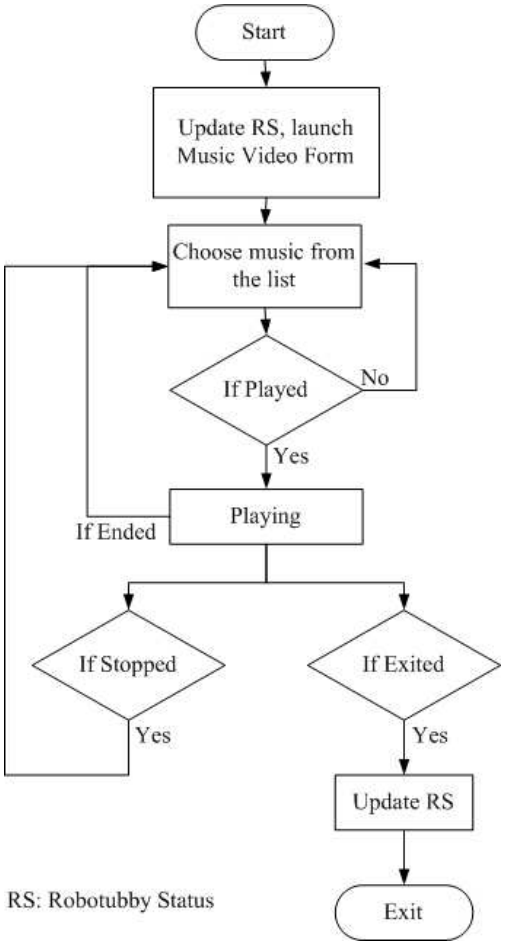


Figure 3.18: The flowchart of playing music video function.

when Robotubby speaks out the given answers, it can move its mouth.

To fulfill this function, we have employed AIMLBot which is a programme implementation of an AIML (Artificial Intelligence Markup Language) and can be directly downloaded from the internet [79]. By using this technique, the user can chat the computer with natural languages.

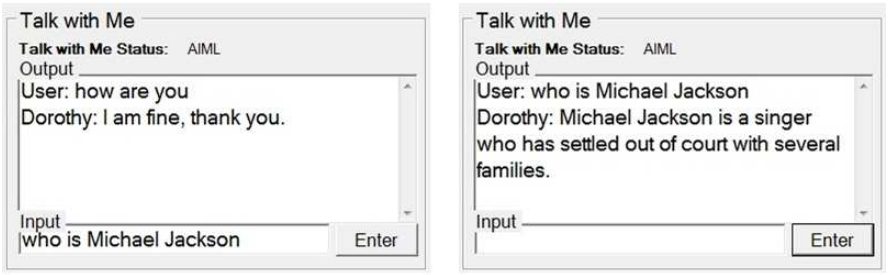


Figure 3.19: The sub-interface of chatting with a child.

Figure 3.20: The sub-interface of video calling.

3.4.7 Video Calling

The function of video calling is developed to build communication between a user like a child and another user like a parent through two computers and the internet. Through video calling, a child can talk with his parent by using voice and video. Figure 3.20 shows the sub-interface of this function which includes the buttons of "Start" and "Exit". After clicking the button of "Start" of both two computers, talking starts. After clicking the button of "Exit" of only one computer, talking ends. We developed this function by taking [80] and [81] as references.

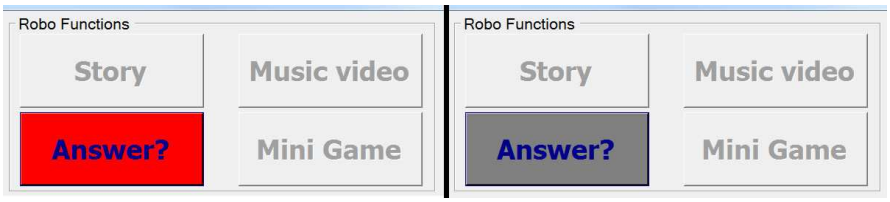


Figure 3.21: The blinking notification button for the incoming call.

Since video calling function aims to be used in two computers that are placed in two different locations, the notification system which includes reminding text, flashing button, and ringing tone has been designed to inform the incoming call. Figure 3.21 illustrates the blinking notification button for the incoming call. When there is an incoming call from the other computer, the text of button will change from "Video Call" to "Answer?" and the background color of button will change from gray to red. At the same time, the ringing tone will ring out. Socket programming in C♯ was employed to transfer information between two computers.

3.5 Summary

In this chaper, we have developed a robotic nanny named Dorothy Robotubby with the aims to play with and take care of a child during his/her parent or caregiver absences. Robotubby is upper-body and about 70cm in height. It is designed with a caricatured appearance, and consists of a head, a neck, a body, two arms, two hands, and a touch screen in its belly. Robotubby includes two interaction interfaces. One interface is developed on the touch screen for a child to control the robot to accomplish several tasks like storytelling, playing music, games, chatting, and video call. When Robotubby executes the first four tasks, it can demonstrate different facial expressions and actions. While for the other interface, it is mainly

employed to remotely control the robot by the child's parent. Meanwhile, the parent can see his/her child through images from Robotubby. When the child operates the robot, two functions including face tracking and emotion recognition will work all the time.

In summary, the developed robotic nanny system can not only interact with a child, but also build a connection between a child and his/her parent.

Chapter 4

Misalignment-Robust Facial Expression Recognition

4.1 Introduction

Automatic facial expression recognition plays an important role in human emotion perception and social interaction, and has therefore attracted much attention in the area of pattern recognition, computer vision, human-computer interaction, and human-robot interaction. Since appearance-based methods are popular and have demonstrated reasonable performance in terms of the recognition accuracy, they have been widely used in many facial expression recognition systems.

It is generally believed that the intrinsic dimensionality of the facial feature space is much lower than that of the original face image space. Hence, it is necessary to apply an efficient and effective feature extraction method to reduce the feature dimensionality of face images for feature representation and recognition. Subspace learning techniques are such methods which can reveal the intrinsic dimensionality of the original images and obtain some succinct and compact features, and hence

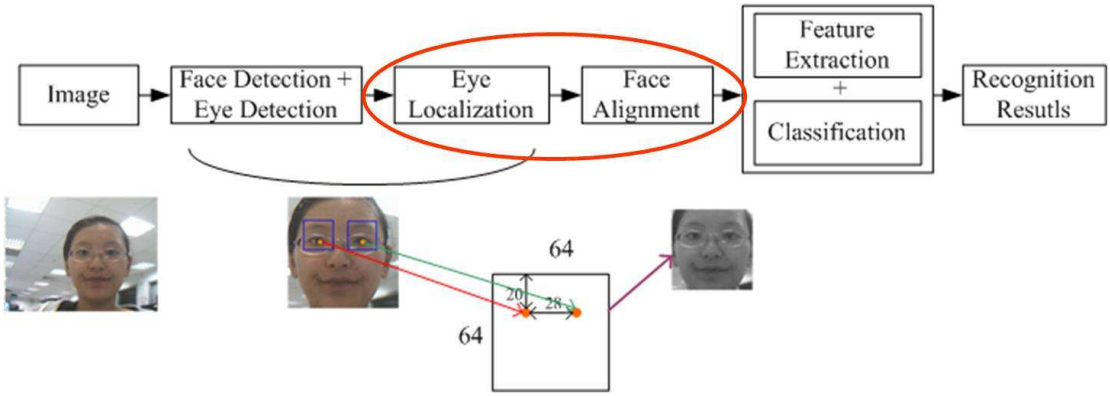


Figure 4.1: The flowchart of an automatic facial expression recognition system.

they have been widely used for facial expression recognition in recent years. By applying these methods, face images are projected into a low-dimensional subspace which is optimal for data reconstruction or recognition.

Figure 4.1 illustrates the flowchart of an automatic facial expression recognition system. This system mainly consists of face and eye detection, eye localization, face alignment, and facial expression recognition. Since facial images for recognition are pre-processed by face alignment, the accuracy of face alignment usually affects the performance of facial expression recognition. Moreover, face alignment depends heavily on the performance of eye localization. Therefore, face alignment templates are used as the eye position baseline. If eye positions are wrongly localized, face misalignment will occur.

Most existing appearance-based facial expression recognition methods, however, can only work well when face images are well-aligned. In many real-world applications such as human-robot interaction and visual surveillance, it is very challenging to obtain well-aligned face images for recognition, especially under uncontrolled conditions. Hence, there are usually some spatial misalignments in

the cropped face images due to the eye localization errors even if the eye positions are manually located. A natural question is how spatial misalignments affect the performance of these appearance-based facial expression recognition methods. While a large number of facial expression recognition methods have been proposed in the literature, there is only a few studies on investigating this problem. In this chapter, we first empirically investigate this problem, and then propose a new misalignment-robust method to extract discriminative features for facial expression recognition. Motivated by the fact that facial images from different expressions (interclass samples) with small differences are more easily mis-classified than those with large differences, we propose a biased linear discriminant analysis (BLDA) method by imposing large penalties on interclass samples with small differences and small penalties on those samples with large differences simultaneously, such that more discriminative features can be extracted for recognition. Moreover, we further propose using the IMage Euclidean Distance (IMED) [82] rather than the widely used Euclidean distance to seek a low-dimensional subspace for facial feature extraction, such that the subspace sought is more discriminative and robust. Experimental results on two widely used face databases are presented to show the efficacy of the proposed method.

The rest of this chapter is organized as follows. Section 4.2 empirically shows how spatial misalignment affects existing appearance-based methods for facial expression recognition. Section 4.3 presents our proposed approach. Section 4.4 presents experimental results, and Section 4.5 concludes the chapter.

4.2 Empirical Study of Appearance-Based Facial Expression Recognition with Spatial Misalignments

4.2.1 Data Sets

Two publicly available facial expression image databases including the Cohn-Kanade [59] and JAFFE [83] databases were selected to conduct facial expression recognition with spatial misalignments experiments to investigate the performance of existing appearance-based methods.

The Cohn-Kanade database consists of 100 university students aged from 18 to 30 years. 65% subjects are female, 15% are African-American, and 3% are Asian or Latino. Subjects are instructed to perform a series of 23 facial displays, seven of which are anger, disgust, fear, happy, neutral, sad, and surprise. We selected 14 subjects which contain all the seven different expressions from the database, where each expression has three samples. Hence, we have 294 samples in total. As the original image sequences in the database start from a neutral expression and end with the peak of the expression, we selected the last three frames of each expression sequence. For the neutral expression, we selected the first frame of three different sequences. The size of the original sample is 640×490. We first manually located the eye positions to obtain the real eye coordinates (x_1, y_1) and (x_2, y_2) of the left and right eyes of each image. Then, we applied two random vectors (l_1, l_2) and (r_1, r_2) to the coordinates of the left and right eyes, respectively,

as follows:

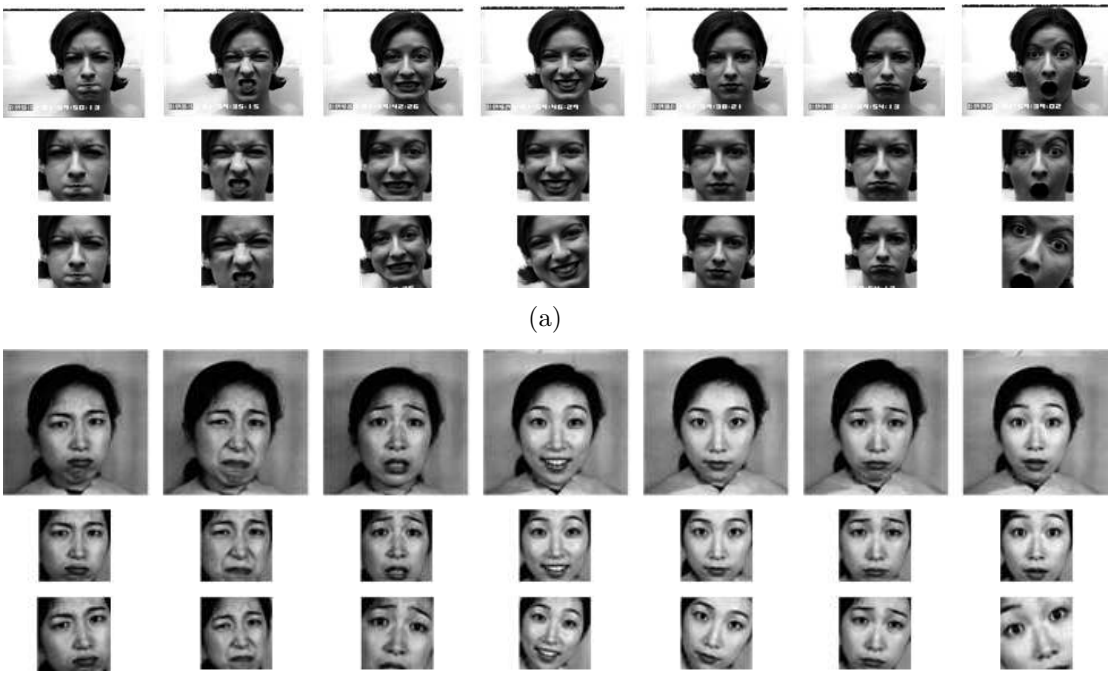
$$\begin{cases} x_1^n = x_1 + l_1 \\ y_1^n = y_1 + l_2 \\ x_2^n = x_2 + r_1 \\ y_2^n = y_2 + r_2 \end{cases} \tag{4.2.1}$$

where (x_1^n, y_1^n) and (x_2^n, y_2^n) are the mis-located eye coordinates , (l_1, l_2) and (r_1, r_2) are uncorrelated and normally distributed with zero mean and a standard deviation of T. We aligned the face images with real and mis-located eye coordinates and resized them into 64×64. Figure 4.2(a) shows some original, well-aligned and misaligned images of different expressions of one subject from the Cohn-Kanade database, which are shown from top to down respectively.

The JAFFE database consists of 213 facial expression images from 10 Japanese females. They posed 3 or 4 examples for each of the seven basic expressions (six emotional expressions including anger, disgust, fear, happy, sad, surprise plus neutral expression). The image size is 256×256. Similar to the Cohn-Kanade database, we manually located the eye positions of the images and aligned and resized them into 64×64 with and without spatial misalignments, and Figure 4.2(b) shows some original, well-aligned and misaligned images of different expressions of one subject from the JAFFE database.

4.2.2 Results

Similar to [61], we also applied four popular subspace learning methods including principal component analysis (PCA) [31], linear discriminant analysis (LDA) [32],



(a)

(b)

Figure 4.2: Examples of the original, well-aligned, and misaligned images of one subject from the (a) Cohn-Kanade and (b) JAFFE databases. From left to right are the facial images with anger, disgust, fear, happy, neutral, sad, and surprise expressions, respectively.

locality preserving projections (LPP) [33] and orthogonal neighborhood preserving projections (ONPP) [34] for appearance-based facial expression recognition with spatial misalignments. We adopted a 10-fold cross-validation strategy in our evaluation: 90% of the samples were used for training and 10% for testing. We chose the nearest neighbor (NN) classifier for recognition. Figures 4.3 and 4.4 show the recognition performance of these four methods versus different amounts of misalignments on the Cohn-Kanade and JAFFE databases, respectively.

We can clearly observe from Figures 4.3 and 4.4 that spatial misalignments indeed

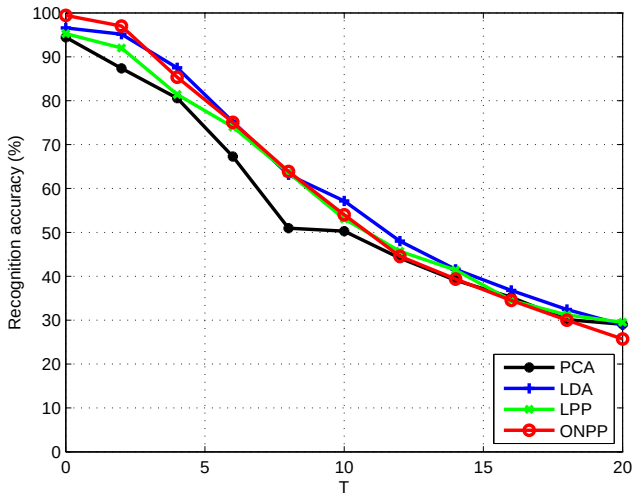


Figure 4.3: Recognition accuracy versus different amounts of spatial misalignments on the Cohn-Kanade database.

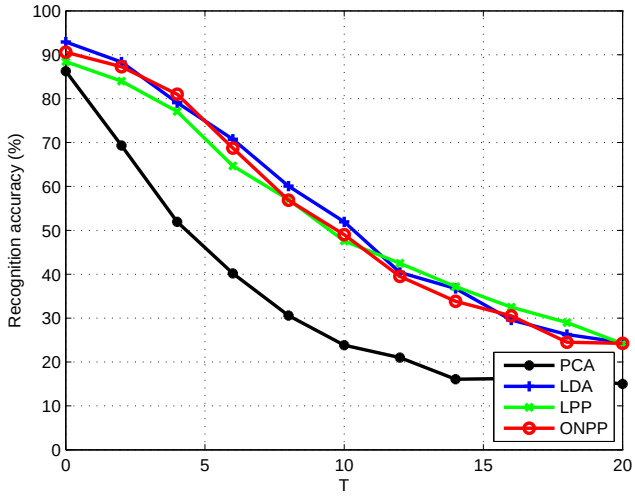


Figure 4.4: Recognition accuracy versus different amounts of spatial misalignments on the JAFFE database.

affect the recognition accuracy of conventional subspace learning-based facial expression recognition methods. Moreover, the larger the spatial misalignment is, the worse the performance is. In many real-world applications, it is still very challenging to precisely localize the eye positions of face images for image alignment, especially under uncontrolled conditions. Hence, it is desirable to develop misalignment-robust methods for facial expression recognition.

4.3 Proposed Approach

Generally speaking, subspace learning techniques can be mainly classified into two categories: supervised-based and unsupervised-based. As supervised learning methods usually outperform unsupervised ones for facial expression recognition tasks and LDA is a popular and widely used supervised subspace learning method due to its simplicity and effectiveness, we employed and modified LDA in this section to learn a new feature space to implement our misalignment-robust facial expression recognition task. We first briefly review LDA in the following.

4.3.1 LDA

Considering a set of facial images denoted as $X = [x_1, x_2, \cdots, x_N]$, $x_i \in R^d$, $i = 1, 2, \cdots, N$, where N is the number of samples and d is the feature dimension of each face sample. For supervised subspace learning algorithms such as LDA, the class label of x_i is assumed to be $l_i \in \{1, 2, \cdots, c\}$, where c is the number of classes. For the jth class, n_j denotes the number of its samples, where $j = 1, 2, \cdots, c$. Hence, $N = \sum_{j=1}^{c} n_j$. Generally, the objective of a subspace learning algorithm

is to find a linear projection matrix $W = [w_1, w_2, \cdots, w_k]$ to map x_i into a low-dimensional representation y_i, where $y_i = W^T x_i$, $y_i \in R^m$, $m \ll d$, and y_i preserves the main information of the original data.

LDA seeks to find a set of projection axes such that the Fisher criterion (the ratio of between-class scatter to within-class scatter) is maximized after the projection. The between-class scatter S_B and the within-class scatter S_W are defined as [32, 84]

$$S_B = \frac{1}{N} \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T \tag{4.3.1}$$

$$S_W = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - m_i)(x_{ij} - m_i)^T \tag{4.3.2}$$

where x_{ij} denotes the jth training sample of the ith class, m_i is the mean of the training samples of the ith class, and m is the mean of all the training samples. The objective of LDA is defined as

$$\max_w \frac{w^T S_B w}{w^T S_W w} \tag{4.3.3}$$

The corresponding projections $\{w_1, w_2, \cdots, w_k\}$ comprise a set of the eigenvectors of the following generalized eigenvalue function

$$S_B w = \lambda S_W w \tag{4.3.4}$$

Let $\{w_1, w_2, \cdots, w_k\}$ be the eigenvectors corresponding to the k largest eigenvalues $\{\lambda_i | i = 1, 2, \cdots, k\}$ ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. Then $W = [w_1, w_2, \cdots, w_k]$ is the LDA projection. Note that the rank of S_B is bounded by $c - 1$ [84, 32], i.e., k is at most equal to $c - 1$.

4.3.2 BLDA

While LDA has attained reasonably good performance in many facial expression recognition applications, there is still one shortcoming in LDA: all classes are equally treated in the procedure of the feature learning. For facial expression recognition, different classes (expressions) may have different similarities, hence, the difficulties to correctly recognize them are undoubtedly different. For example, for a testing sample with sad expression, it is much easier to mis-recognize it as disgust rather than happy expression. Motivated by this observation, we propose here a new biased linear discriminant analysis (BLDA) method by imposing large penalties on interclass samples with small differences and small penalties on those samples with large differences simultaneously, such that more discriminative features can be extracted for recognition. Specifically, we formulate BLDA into the following optimization objective:

$$\max_{w} \frac{w^T \hat{S_B} w}{w^T \hat{S_W} w} \tag{4.3.5}$$

where

$$\hat{S_B} = \sum_{i=1}^{c} \sum_{j=1}^{c} g(i,j)(m_i - m_j)(m_i - m_j)^T \tag{4.3.6}$$

$$\hat{S_W} = \sum_{k=1}^{c} \sum_{x_i \in c_k} (x_i - m_k)(x_i - m_k)^T \tag{4.3.7}$$

$g(i,j)$ is a penalty function to impose different weights to characterize the relationship between the ith and jth classes in calculation of the between-class scatter $\hat{S_B}$. As discussed before, the larger the similarities are, the higher penalty should be imposed, and the higher $g(i,j)$ should be assigned. Obviously, there are a number of potential strategies to define the penalty function $g(i,j)$, and it is generally

believed that $g(i,j)$ can be a monotone function of the distance between the ith and the jth classes.

Let $d_{i,j}$ be the distance between m_i and m_j, we define $g(i,j)$ as follows:

$$g(i,j) = d_{i,j}^{\alpha} \tag{4.3.8}$$

where

$$d_{i,j} = \frac{< m_i, m_j >}{\|m_i\|_2 \cdot \|m_j\|_2} \tag{4.3.9}$$

$\|x\|_2$ denotes the L_2 norm of x, and $\alpha \geq 0$.

There are two reasons for us to apply the correlation metric rather than the Euclidean distance to measure the similarity. On one hand, facial images often need to be preprocessed such as normalization and histogram equalization, and correlation is more robust than Euclidean distance against such nonlinear changes of the data distribution. On the other hand, many studies have shown that correlation metric-based similarity measurement outperforms the conventional Euclidean distance for the classification task [85].

Having obtained $\hat{S_B}$ and $\hat{S_W}$, the feature space of BLDA comprises a set of the eigenvectors of the generalized eigenvalue function $\hat{S_B}w = \eta\hat{S_W}w$. Similar to LDA, let $\{w_1, w_2, \cdots, w_k\}$ be the eigenvectors corresponding to the k largest eigenvalues $\{\eta_i | i = 1, 2, \cdots, k\}$ ordered such that $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_k$. Then $W = [w_1, w_2, \cdots, w_k]$ is the BLDA projection. It should also be noted that original LDA method is the special case of our proposed BLDA method when $\alpha = 0$.

To show the advantages of the BLDA, we selected 10 subjects from the Cohn-Kanade database, each subject contains 7 expressions and each expression has 3

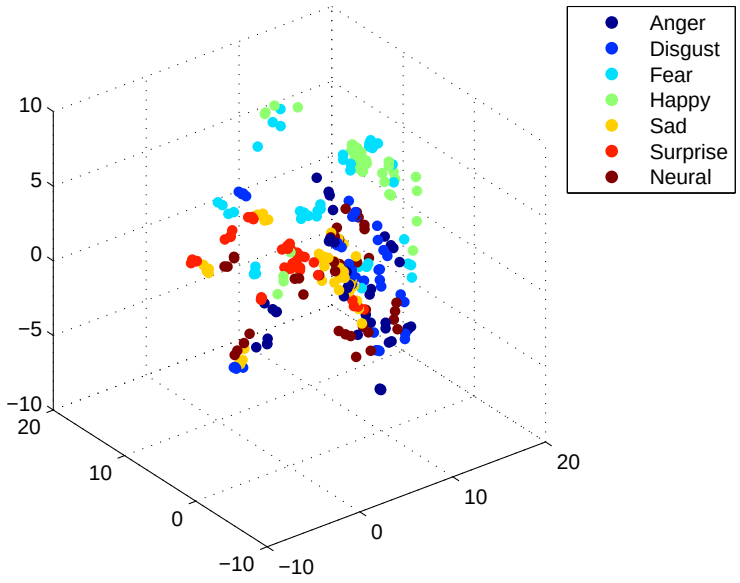


Figure 4.5: The projections of the first three components of the original data on the PCA feature space.

samples. We visualized their original distribution in Figure 4.5. Note that for ease of presentation, we only used a three-dimensional space by PCA. Figures 4.6 and 4.7 show the low-dimensional distributions in the conventional LDA and our proposed BLDA subspaces, respectively. We can see that compared with LDA, BLDA can better separate interclass samples, especially those interclass samples with high similarity, such as the sad and disgust samples, which further shows that more discriminative information can be revealed in the learned BLDA subspace.

4.3.3 IMED-BLDA

Most existing subspace analysis methods learn a low-dimensional feature subspace by using the Euclidean metric, however, they usually suffer from a high sensitivity

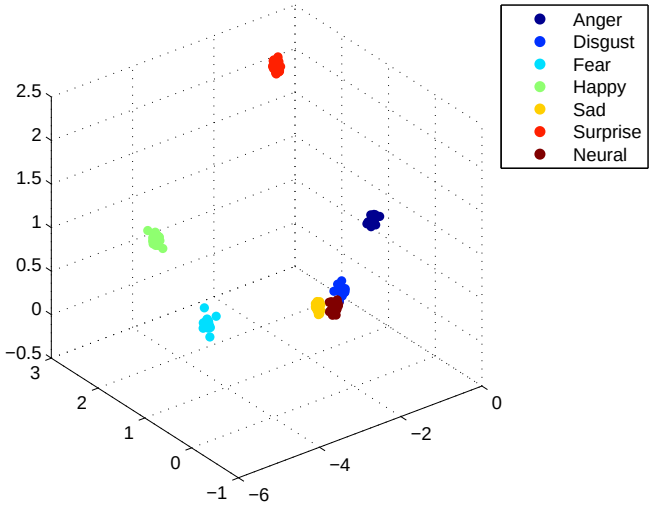


Figure 4.6: The projections of the first three components of the original data on the LDA feature space.

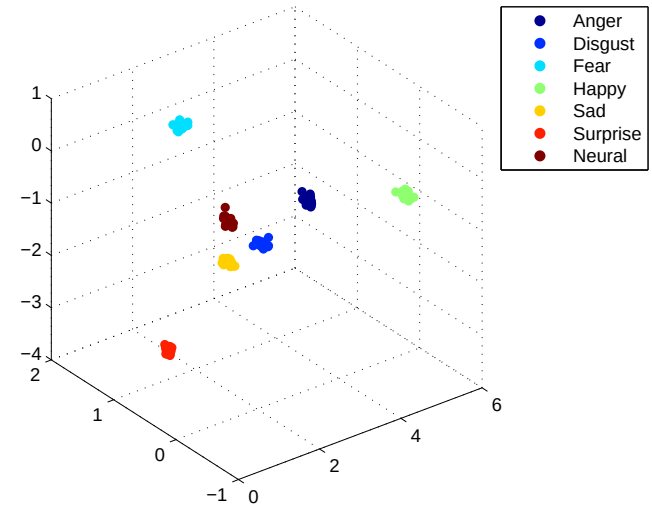


Figure 4.7: The projections of the first three components of the original data on the BLDA feature space. Note that here α is set to be 50 for BLDA. For interpretation of color in this figure, please refer to the original enlarged color pdf file.

to a small deformation because the Euclidean metric does not take into account the spatial relationship and a small spatial misalignment may result in a large Euclidean distance. To address this problem, Wang *et al.* [82] proposed an IMage Euclidean Distance (IMED) to better characterize the dissimilarity of two samples when small deformations are involved.

Let $x = [x_1, x_2, \cdots, x_{MN}]$ and $y = [y_1, y_2, \cdots, y_{MN}]$ be two $M \times N$ images, where x_i and y_i are pixels of these two images, respectively. The IMED between x and y is defined as

$$\begin{aligned} d_{i,j}^{IMED} &= \sqrt{\sum_{i=1}^{MN}\sum_{j=1}^{MN} g_{ij}(x_i - y_i)(x_j - y_j)} \\ &= (x_i - y_i)^T G (x_j - y_j) \end{aligned} \tag{4.3.10}$$

where the symmetric and positive define matrix G is referred to a metric matrix, and g_{ij} is the metric coefficient indicating the spatial relationship between pixels x_i and y_j. The definition of g_{ij} is given by

$$g_{ij} = f(d_{ij}^s) = \frac{1}{2\pi\sigma^2} \exp(-\frac{d_{ij}^s}{2\sigma^2}) \tag{4.3.11}$$

where d_{ij}^s is the spatial distance between x_i and y_j on the image lattice, and σ is the width parameter. For example, if x_i is at location (t_1, t_2) and y_j is at location (t_1', t_2'), then d_{ij}^s is calculated as

$$d_{ij}^s = \sqrt{(t_1 - t_1')^2 + (t_2 - t_2')^2} \tag{4.3.12}$$

Now, we use the IMED metric instead of the Euclidean metric to learn a new feature subspace, called IMED-BLDA, by modifying the within-class and between

class scatters as follows:

$$\hat{S_B}^{IMED} = \sum_{i=1}^{c}\sum_{j=1}^{c} g(i,j)d^{IMED}(m_i, m_j) \tag{4.3.13}$$

$$\hat{S_W}^{IMED} = \sum_{k=1}^{c}\sum_{x_i \in c_k} d^{IMED}(x_i, c_k) \tag{4.3.14}$$

Similar to BLDA, the projections of IMED-BLDA $\{w_1^{IMED}, w_2^{IMED}, \cdots, w_k^{IMED}\}$ comprise a set of the eigenvectors of the following generalized eigenvalue function

$$\hat{S_B}^{IMED} w^{IMED} = \lambda \hat{S_W}^{IMED} w^{IMED} \tag{4.3.15}$$

Let $\{w_1^{IMED}, w_2^{IMED}, \cdots, w_k^{IMED}\}$ be the eigenvectors corresponding to the k largest eigenvalues $\{\lambda_i | i = 1, 2, \cdots, k\}$ ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. Then $W^{IMED} = [w_1^{IMED}, w_2^{IMED}, \cdots, w_k^{IMED}]$ is the IMED-BLDA projections.

Since IMED considers the spatial relationship between pixels, it is expected to be robust to spatial misalignments. To verify this point, Figure 4.8 plots the trace ratio of the between-class scatter to within-class scatter of BLDA on the Cohn-Kanade database by using the Euclidean and IMED distances versus different amounts of spatial misalignments, respectively. Generally, the larger the ratio is, the more the separability of the subspace is. We can observe from this figure that IMED is better than the Euclidean distance in characterizing this ratio. Since the trace ratio is closely related to the recognition accuracy, we expect the IMED metric used in BLDA can achieve higher recognition accuracy. We will show the recognition accuracy in Section 4.4.

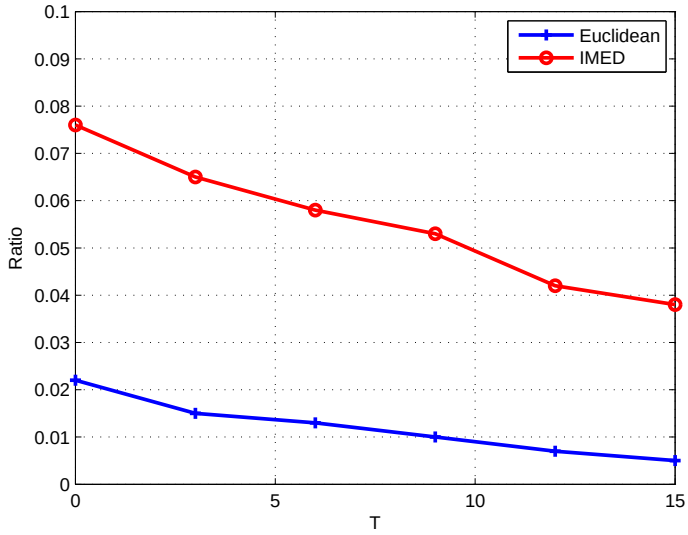


Figure 4.8: The ratio of the trace of the between-class scatter to the trace of the within-class scatter by using the Euclidean and IMED distances on the Cohn-Kanade database. It is easy to observe from this figure that IMED is better than the Euclidean distance in characterizing this ratio. Moreover, the larger amounts of the misalignment, the better performance obtained.

4.4 Experimental Results

We conducted facial expression recognition experiments on the Cohn-Kanade and JAFFE databases. The data used here are the same as those used in Section 4.2. Similarly, the 10-fold cross-validation strategy and the NN classifier are employed for recognition. We also compared the proposed IMED-BLDA method with the most effective conventional subspace learning methods including LDA and LPP. The reason we selected LDA and LPP for comparison here is that LDA and LPP can be performed in a supervised setting, PCA and ONPP are usually unsupervised, and supervised methods generally outperform unsupervised ones for classification tasks. To provide a fair comparison, all the results reported here are

Table 4.1: Recognition performance comparison on the Cohn-Kanade database.

Method	$T = 0$	$T = 3$	$T = 6$	$T = 9$	$T = 12$	$T = 15$
LDA	96.36	91.07	74.29	55.71	45.00	38.57
LPP	92.86	84.29	70.71	57.14	46.79	36.07
IMED-LDA	98.21	92.14	77.14	63.21	48.57	41.07
BLDA	97.36	91.67	76.29	62.87	48.23	39.65
IMED-BLDA	**98.51**	**92.64**	**78.24**	**65.51**	**50.57**	**44.07**

Table 4.2: Recognition performance comparison on the JAFFE database.

Method	$T = 0$	$T = 3$	$T = 6$	$T = 9$	$T = 12$	$T = 15$
LDA	92.92	86.33	70.75	58.08	40.42	35.34
LPP	88.42	83.56	64.67	52.34	38.86	33.54
IMED-LDA	93.23	87.54	71.24	60.22	41.46	36.44
BLDA	93.22	88.02	72.75	61.08	42.42	37.24
IMED-BLDA	**94.56**	**89.34**	**73.45**	**62.34**	**43.68**	**39.44**

based on the best tuned parameters of all the compared methods. Specifically, α is empirically set to be 50 for IMED-BLDA in our experiments.

As the advantage of the proposed IMED-BLDA approach stems from two different aspects: the IMED metric and the weighted function, we also evaluated the performance when only one factor is applied to reveal their respective effects. We thus formulated two other LDA-based algorithms, i.e., BLDA and IMED-LDA. We reported here the best result of each algorithm under comparison by exploring all possible feature dimensions. The average recognition accuracies are tabulated in Tables 4.1 and 4.2. As can be seen, the proposed IMED-BLDA method always outperforms the other compared methods in terms of recognition accuracy.

To better show the effectiveness of the IMED metric for misalignment-robust facial expression recognition, we used this metric to other three subspace analysis methods: PCA, LPP and ONPP to formulate the corresponding IMED-based methods.

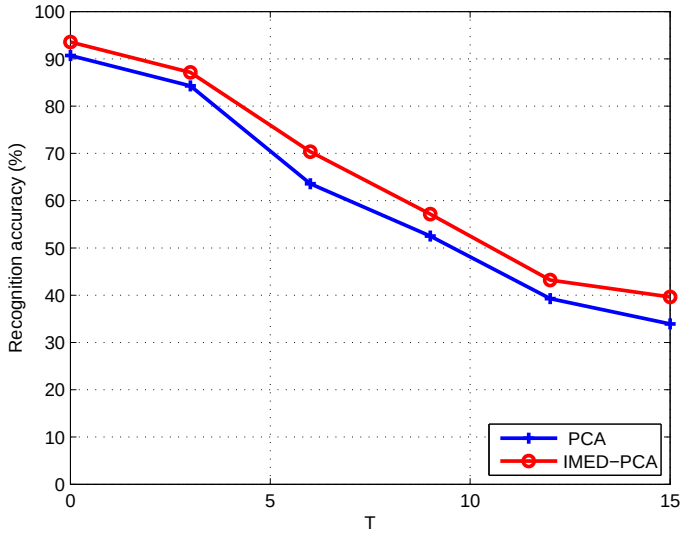


Figure 4.9: Performance comparisons of PCA and IMED-PCA subspace methods learned by the Euclidean and IMED metric, respectively.

Figures 4.9-4.11 show the recognition performance of these methods versus different amounts of misalignments on the Cohn-Kanade database. We can easily observe from these three figures that our proposed IMED-based subspace methods consistently outperform existing Euclidean-based subspace learning methods for facial expression recognition with spatial misalignments, which further demonstrates the effectiveness of the proposed approach.

Lastly, we evaluated the robustness of the IMED-BLDA versus different values of the parameter α, and plotted the recognition accuracy in Figure 4.12. We can see from this figure that IMED-BLDA is robust and can achieve good performance in a large range of α. Hence, it is easy to set an appropriate value of α for practical applications.

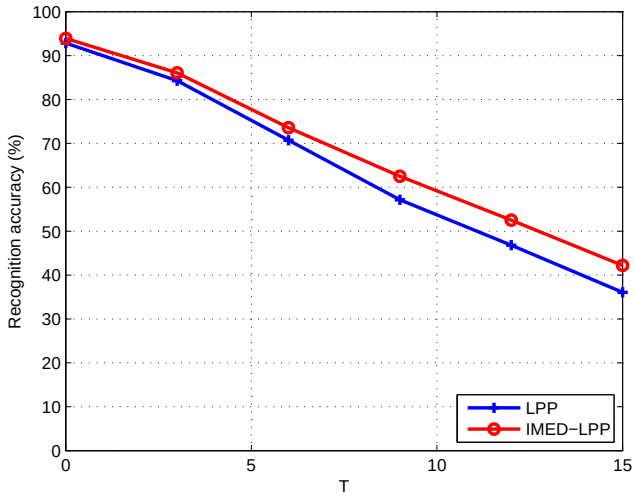


Figure 4.10: Performance comparisons of LPP and IMED-LPP subspace methods learned by the Euclidean and IMED metric, respectively.

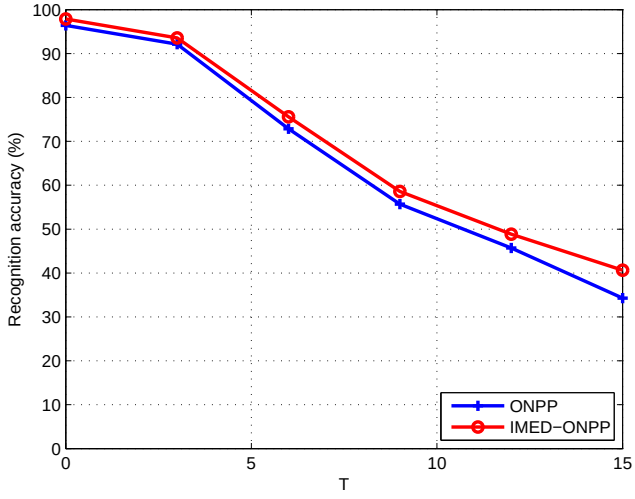


Figure 4.11: Performance comparisons of ONPP and IMED-ONPP subspace methods learned by the Euclidean and IMED metric, respectively.

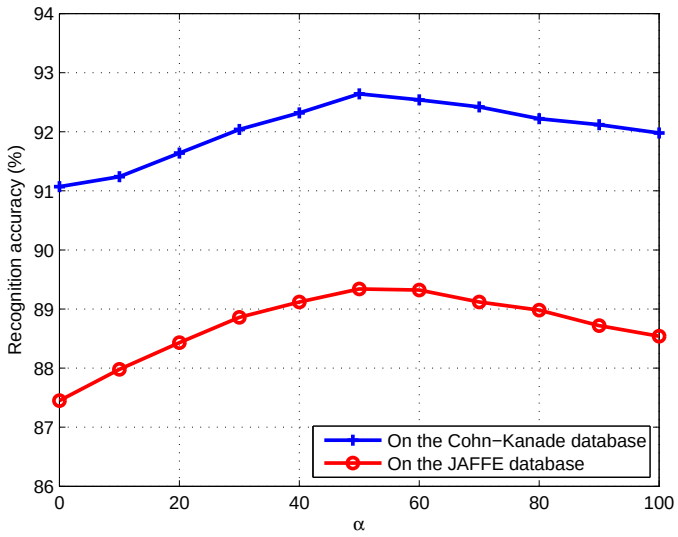


Figure 4.12: The performance of IMED-BLDA versus different values of α.

4.5 Summary

We have proposed in this chapter a new misalignment-robust subspace analysis approach for facial expression recognition. We first empirically showed that spatial misalignments indeed affect the recognition accuracy of conventional subspace learning-based facial expression recognition methods. To make better use of the different interclass samples in learning the feature subspace, we proposed a biased method by imposing large penalties on interclass samples with small differences and small penalties on those samples with large differences simultaneously, such that more discriminative features can be extracted for recognition. Moreover, we learned a robust feature subspace by using the IMage Euclidean Distance (IMED) rather than the widely used Euclidean distance, such that the subspace sought is more discriminative and robust to spatial misalignments. Experimental results

on two widely used face databases have demonstrated the efficacy of the proposed methods.

For future work, we want to further extend the proposed misalignment-robust subspace analysis approach to other supervised manifold learning methods to explore the nonlinear manifold structure of facial expression data. Moreover, how to design a better penalty function to further improve the recognition performance remains another interesting direction of future work. We are also going to collect more facial expression images under uncontrolled environments to examine the robustness of our proposed method in real-world applications. In this study, we only assume there is spatial misalignment in facial images, however, this assumption may not hold because there could be some other variations in facial expression images such as varying illumination, poses, and occlusions, even for the same person. Hence, how to simultaneously deal with the spatial misalignment as well as other variations for robust facial expression recognition remains more investigation in the future.

Chapter 5

Cross-Dataset Facial Expression Recognition

5.1 Introduction

Appearance-based techniques have been widely used to reveal the intrinsic structure of data and applied for facial expression recognition. By using these methods, facial expression images are projected into a low-dimensional feature space to reduce the feature dimensions. Representative and state-of-the-art methods include principal component analysis (PCA) [31], linear discriminant analysis (LDA) [32], locality preserving projections (LPP) [33] and orthogonal neighborhood preserving projections (ONPP) [34]. Recently, Shan *et al.* [29] compared these methods for facial expression recognition and reported that supervised LPP was the best one in supervised methods and ONPP produced the best results in unsupervised methods.

Most existing facial expression recognition methods assume facial images in the training and testing sets are collected under the same condition such that they

are independent and identically distributed. However, in many real-world applications, this assumption may not hold as the testing data are usually collected online and generally more uncontrollable than the training data, such as different races, illuminations, and imaging conditions. Under this scenario, the performance of conventional subspace learning methods may be poor because the training and testing data are not independent and identically distributed which refers to cross-dataset problem. The generalization capability of these methods is limited on the cross-dataset facial expression recognition problem. To the best of our knowledge, this problem is seldom addressed in the literature for facial expression recognition even if it is very important to drive facial expression recognition into real applications.

To address this problem, we propose a new transfer subspace learning approach to learn a feature space which transfers the knowledge gained from the training set to the target (testing) data to improve the recognition performance under cross-dataset scenarios. We apply the proposed approach to four popular subspace learning methods including PCA, LDA, LPP and ONPP, and formulate the corresponding transfer PCA (TPCA), transfer LDA (TLDA), transfer LPP (TLPP) and transfer ONPP (TONPP) for cross-dataset facial expression recognition. Experimental results are presented to demonstrate the efficacy of the proposed approaches.

The rest of this chapter is organized as follows. Section 5.2 reviews some related work on subspace learning and transfer learning. Section 5.3 presents our proposed methods. Section 5.4 presents experimental results to demonstrate the effectiveness of the proposed methods, and Section 5.5 concludes the chapter.

5.2 Related Work

5.2.1 Subspace Learning

Let $X = [x_1, x_2, \cdots, x_N]$, $x_i \in R^d$, $i = 1, 2, \cdots, N$, be a training set of facial images, where N is the number of samples and d is the feature dimension of each sample. For supervised subspace learning algorithms, the class label of x_i is assumed to be $l_i \in \{1, 2, \cdots, c\}$, where c is the number of classes. For the jth class, n_j denotes the number of its samples, where $j = 1, 2, \cdots, c$. Hence, $N = \sum_{j=1}^{c} n_j$. The objective of a subspace learning algorithm, such as PCA, LDA, LPP and ONPP, is to find a linear projection matrix $W = [w_1, w_2, \cdots, w_k]$ to map x_i into a low dimensional representation y_i, where $y_i = W^T x_i \in R^m$, $m < d$ [29]. The essential differences of different subspace learning methods lie in their differences in defining and finding the projection matrix W by using different objective functions and constraints, such as

$$\begin{aligned} \min \quad & F(W) \\ \text{subject to} \quad & G(W) = 0 \end{aligned} \tag{5.2.1}$$

Table 5.1 shows the objective functions and constraints of PCA, LDA, LPP and ONPP, where $S_T = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^T$, $m = \frac{1}{N}\sum_{i=1}^{N} x_i$, $S_B = \frac{1}{N}\sum_{i=1}^{c} n_i(m_i - m)(m_i - m)^T$, $S_W = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_{ij} - m_i)(x_{ij} - m_i)^T$, x_{ij} denotes the jth training sample of the ith class, m_i is the mean of the training samples of the ith class, $L = D - S$, $D_{ii} = \sum_j S_{ji}$, S_{ij} is the locality similarity between x_i and x_j, $M =$

Table 5.1: Objective functions and constraints of four popular subspace learning methods.

Method	$F(W)$	$G(W)$
PCA	$-tr(W^T S_T W)$	$W^T W - I = 0$
LDA	$\frac{tr(W^T S_W W)}{tr(W^T S_B W)}$	$-$
LPP	$W^T X L X^T W$	$W^T X D X^T W - I = 0$
ONPP	$tr(W^T X M X^T W)$	$W^T W - I = 0$

$(I - V^T)(I - V)$, V can be obtained by solving the following optimization function:

$$\min \varepsilon(V) = \sum_i \|x_i - \sum_k V_{ik} x_{ik}\|^2 \tag{5.2.2}$$

where x_{ik} is the k-nearest neighbor of x_i.

5.2.2 Transfer Learning

The past five years have witnessed the significance of transfer learning for practical applications such as cross-domain image and text classification, and domain-adaptation video analysis. Transfer learning has also been identified to be an effective solution to address the cross-dataset recognition problem because it can transfer the knowledge gained from the training set to the testing set. Generally, there are three main issues in transfer learning: what to transfer, how to transfer, and when to transfer. Compared with the conventional machine learning techniques, transfer learning can be mainly classified into three categories: inductive transfer learning, transductive transfer learning, and unsupervised transfer learning. Please refer to [86] for more details.

While a number of transfer learning methods have been proposed recently, there is little effort on transfer learning made for subspace learning. To our knowledge,

Si *et al.* [66] first applied transfer learning techniques to subspace learning by minimizing the distribution distance between the source and target domains in subspace learning algorithms. More recently, Su *et al.* [67] employed the mixture Gaussian model to model the distributions of the data in the source and target domains to make it more consistent with the original LDA method. However, these methods estimate the distribution based on the kernel density estimation (KDE) method and Gaussian model, respectively, which may fail when there is a limited number of samples in the source and target domains. In this chapter, we propose a new nonparametric transfer learning approach to learn a feature space which transfers the knowledge gained from the training set to the target (testing) data to improve the facial expression recognition performance under cross-dataset scenarios.

5.3 Proposed Methods

5.3.1 Basic Idea

Since the training and testing samples are implicitly assumed to be independent and identical distribution, conventional subspace learning algorithms seek a feature subspace W by solving an optimization objective function $F(W)$ and then apply W for feature extraction. As mentioned before, this assumption will not hold for cross-dataset facial recognition problem. Under cross-dataset scenarios, we also need to minimize the difference between the training and testing sets besides optimizing $F(W)$.

Given N_1 training samples $X = [x_1, x_2, \cdots, x_{N_1}]$ and N_2 testing samples $Y =$

$[y_1, y_2, \cdots, y_{N_2}]$, our objective now is seeking a feature space W to optimize $F(W)$ in the training set and minimize the differences between X and Y in W simultaneously. Specifically, we formulate our objective into the following optimization function:

$$\min_W F(W) + \lambda H(W) \tag{5.3.1}$$

where $\lambda \geq 0$ is a parameter to balance the contributions of $H(W)$ and $F(W)$. When $\lambda = 0$, Eq 5.3.1 refers to conventional subspace learning methods.

$H(W)$ is formulated as

$$H(W) = \sum_{i=1}^{N_1} \|W^T x_i - W^T \sum_{j=1}^{k} t_{ij} y_{ij}\|^2 \tag{5.3.2}$$

$y_{i1}, y_{i2}, \cdots, y_{ik}$ are the k-nearest neighbors of x_i, $t_{i1}, t_{i2}, \cdots, t_{ik}$ are the corresponding coefficients, and they can be obtained similarly to the coefficients obtained in the locally linear embedding (LLE) method in [87]. With the help of $H(W)$, we can reconstruct each training sample by using several testing samples, which means the knowledge from the training data can be transferred to the testing data.

We simplify $H(W)$ to the following form

$$\begin{aligned} H(W) &= \sum_{i=1}^{N_1} tr[W^T (x_i - \sum_{j=1}^{k} t_{ij} y_{ij})(x_i - \sum_{j=1}^{k} t_{ij} y_{ij})^T W] \\ &= tr[W^T \sum_{i=1}^{N_1} (x_i - \sum_{j=1}^{k} t_{ij} y_{ij})(x_i - \sum_{j=1}^{k} t_{ij} y_{ij})^T W] \\ &= tr(W^T G W) \end{aligned} \tag{5.3.3}$$

where $G \triangleq \sum_{i=1}^{N_1} [(x_i - \sum_{j=1}^{k} t_{ij} y_{ij})(x_i - \sum_{j=1}^{k} t_{ij} y_{ij})^T]$.

The derivative of $H(W)$ is

$$\frac{\partial H(W)}{\partial W} = 2GW \tag{5.3.4}$$

As different subspace learning methods have different $F(W)$, we include different $F(W)$ for different subspace learning methods and formulate the corresponding transferred ones in the following.

5.3.2 TPCA

From Table 5.1, we can obtain $F(W) = -tr(W^T S_T W)$ for PCA. To make the minimization problem with respect to W well-posed, we impose an orthogonal constraint $W^T W = I$ and formulate TPCA as the following constrained optimization problem:

$$\begin{aligned} \min_{W} \quad & T(W) = -tr(W^T S_T W) + \lambda tr(W^T G W) \\ s.t. \quad & W^T W = I. \end{aligned} \tag{5.3.5}$$

Let $\frac{\partial T(W)}{\partial W} = 0$, we can obtain the projections of TPCA by solving the following eigenvalue equation:

$$(\lambda G - S_T)w = \alpha w \tag{5.3.6}$$

Let $\{w_1, w_2, \cdots, w_p\}$ be the eigenvectors corresponding to the p smallest eigenvalues $\{\alpha_i | i = 1, 2, \cdots, p\}$ ordered such that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_p$. Then $W = [w_1, w_2, \cdots, w_p]$ is the subspace projection of TPCA.

5.3.3 TLDA

From Table 5.1, we can obtain $F(W) = \frac{tr(W^T S_W W)}{tr(W^T S_B W)}$ for LDA. Hence,

$$\frac{\partial F(W)}{\partial W} = 2p_1^{-1} S_W W - 2p_1^{-2} p_2 S_B W \tag{5.3.7}$$

where $p_1 = tr(W^T S_B W)$ and $p_2 = tr(W^T S_W W)$.

As Eq. 5.3.7 is nonlinear and it is nontrivial to derive its closed-form global optimal solution, we modify the trace ratio of LDA to the difference form and seek a global solution by the following optimization problem:

$$\begin{aligned} \min_W T(W) &= tr(W^T(S_W - S_B)W) + \lambda tr(W^T G W) \\ s.t. \qquad & W^T W = I. \end{aligned} \tag{5.3.8}$$

Let $\frac{\partial T(W)}{\partial W} = 0$, we can obtain the projections of TLDA by solving the following eigenvalue equation

$$(\lambda G + S_W - S_B)w = \alpha w \tag{5.3.9}$$

We can obtain the projections of TLDA similarly to that of TPCA.

5.3.4 TLPP

For LPP, $F(W) = W^T X L X^T W$. Hence, TLPP can be formulated as the following constrained optimization problem:

$$\begin{aligned} \min_W \quad & T(W) = W^T X L X^T W + \lambda tr(W^T G W) \\ s.t. \quad & W^T W = I. \end{aligned} \tag{5.3.10}$$

Let $\frac{\partial T(W)}{\partial W} = 0$, we can obtain the projections of TLPP by solving the following eigenvalue equation:

$$(XLX^T + \lambda G)w = \alpha w \tag{5.3.11}$$

We can obtain the projections of TLPP similarly to that of TPCA.

5.3.5 TONPP

For ONPP, $F(W) = tr(W^T XMX^T W)$. Hence, TONPP can be formulated as the following constrained optimization problem:

$$\begin{aligned} \min_W T(W) &= tr(W^T XMX^T W) + \lambda tr(W^T GW) \\ s.t. \qquad & W^T W = I. \end{aligned} \tag{5.3.12}$$

Let $\frac{\partial T(W)}{\partial W} = 0$, we can obtain the projections of TONPP by solving the following eigenvalue equation:

$$(XMX^T + \lambda G)w = \alpha w \tag{5.3.13}$$

We can obtain the projections of TONPP similarly to that of TPCA.

5.4 Experimental Results

5.4.1 Data Preparation

Three publicly available facial expression image databases including the JAFFE [88, 83], Cohn-Kanade [59], and Feedtum [89] databases were selected to evaluate the

effectiveness of the proposed methods for cross-dataset facial expression recognition.

The JAFFE database consists of 213 facial expression images from 10 Japanese females. They posed 3 or 4 examples for each of the seven basic expressions (six emotional expressions including anger, disgust, fear, happy, sad, surprise plus neutral expression). The image size is 256×256.

The Cohn-Kanade database consists of 100 university students aged from 18 to 30 years. 65% subjects are female, 15% are African-American, and 3% are Asian or Latino. Subjects are instructed to perform a series of 23 facial displays, seven of which are anger, disgust, fear, happy, neutral, sad and surprise. We selected 10 subjects which contain all the seven different expressions from the database, where each expression has four samples. Hence, we have 280 samples in total. As the original image sequences in the database start from a neutral expression and end with the peak of the expression, we selected the last four frames of each expression sequence. For the neutral expression, we selected the first frame of four different sequences. The size of the original facial image is 640×490.

The Feedtum database, also known as the FG-NET database, is much more challenging because in the database subjects perform the expressions spontaneously and some of the resulting expressions are not well distinguishable. It contains a set of facial image sequences that show a number of subjects performing the seven different universal expressions defined by Ekman and Friesen. All seven expressions were performed three times by each subject. Since these images were captured under natural circumstances, there could be head movement in the images. In

Figure 5.1: Facial expression images of one subject from the (a) JAFFE, (b) Cohn-Kanade, and (c) Feedtum databases. From left to right are the images with anger, disgust, fear, happy, sad, surprise and neutral expressions, respectively.

order to simplify our experiments, only the images which include frontal faces without large head movement were chosen. We selected 10 subjects which contain all the seven different expressions from the database, where each expression has four samples. Hence, we have 280 samples in total. The size of the original facial image is 320×240.

For all the three databases, we converted the images to gray scale and manually located the eye positions. We cropped the face regions from original images according to the eye positions and resized them to 64×64. No further registration such as alignment of mouth was performed in our experiments. Some examples of the aligned images from the databases are shown in Figure 5.1, where (a), (b) and (c) are the example samples of the JAFFE, Cohn-Kanade and Feedtum databases, respectively.

Based on the three databases, we conducted six sets for cross-dataset facial expression recognition as follows:

1. J2C: the training set is JAFFE and the testing set is Cohn-Kanade;
2. J2F: the training set is JAFFE and the testing set is Feedtum;
3. C2J: the training set is Cohn-Kanade and the testing set is JAFFE;
4. C2F: the training set is Cohn-Kanade and the testing set is Feedtum;
5. F2J: the training set is Feedtum and the testing set is JAFFE;
6. F2C: the training set is Feedtum and the testing set is Cohn-Kanade.

5.4.2 Results

We employed the nearest neighbor (NN) classifier for facial expression recognition. The value of λ was empirically set to be 10 for all the four transfer subspace learning methods. We compared our proposed transfer subspace learning methods with four existing non-transferred subspace learning methods including PCA, LDA, LPP and ONPP for cross-dataset facial expression recognition. Figures 5.2-5.7 show the recognition performance of these methods versus different feature dimensions.

We can easily observe from these figures that our proposed transfer learning methods consistently outperform the conventional subspace learning methods in terms of the recognition accuracy. That is because conventional subspace learning algorithms such as PCA, LDA, LPP and ONPP assume that the training and testing

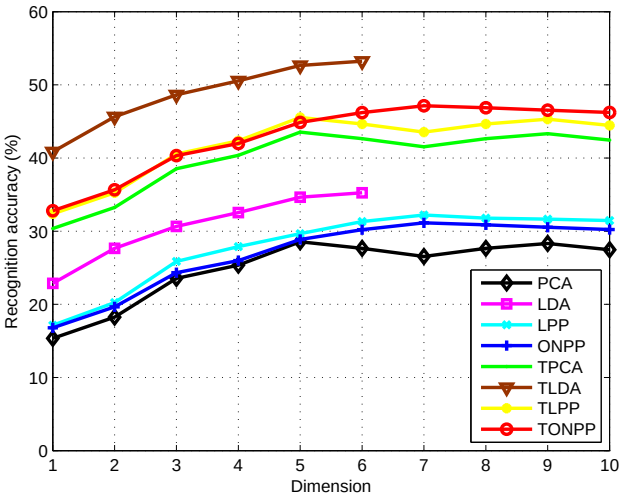


Figure 5.2: Recognition accuracy versus different feature dimensions under the J2C experimental setting.

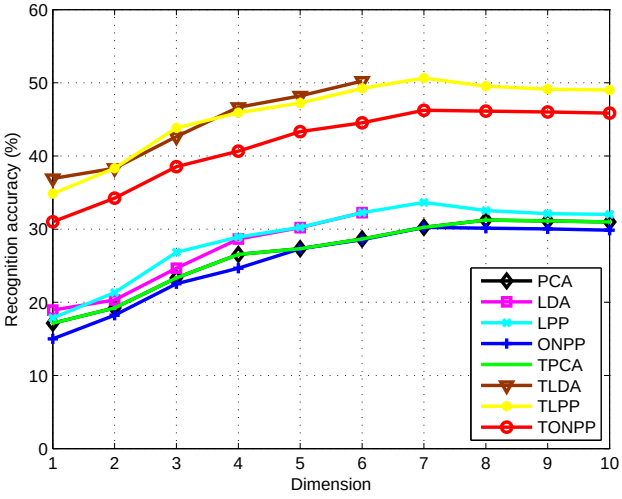


Figure 5.3: Recognition accuracy versus different feature dimensions under the J2F experimental setting.

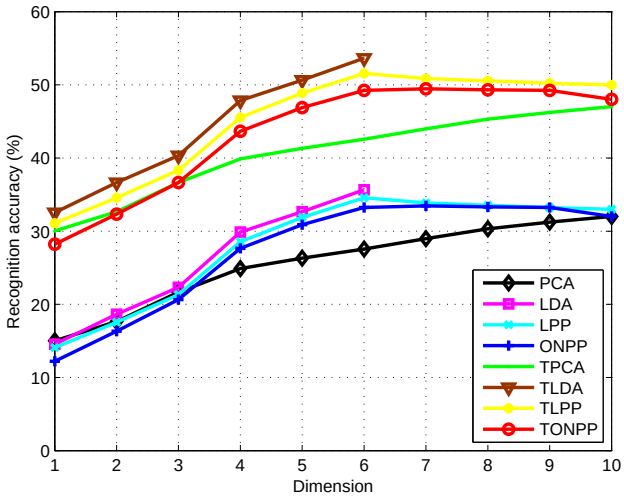


Figure 5.4: Recognition accuracy versus different feature dimensions under the C2J experimental setting.

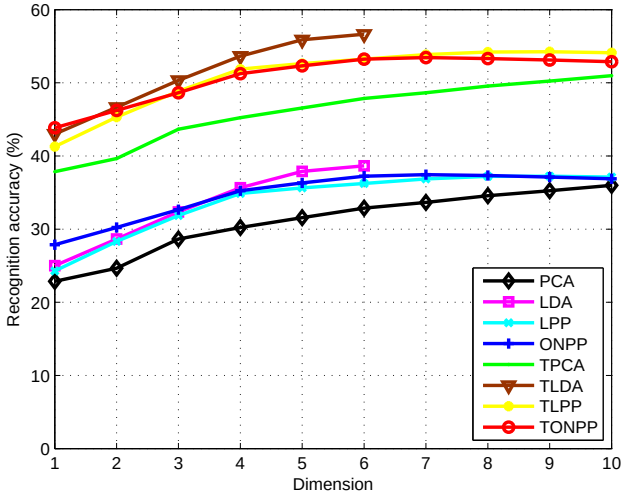


Figure 5.5: Recognition accuracy versus different feature dimensions under the C2F experimental setting.

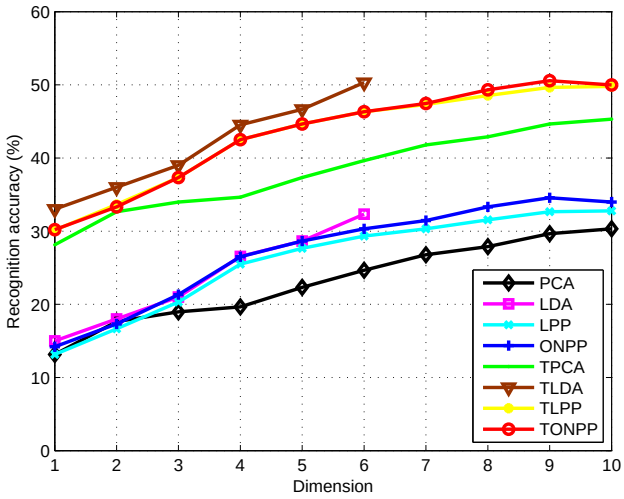


Figure 5.6: Recognition accuracy versus different feature dimensions under the F2J experimental setting.

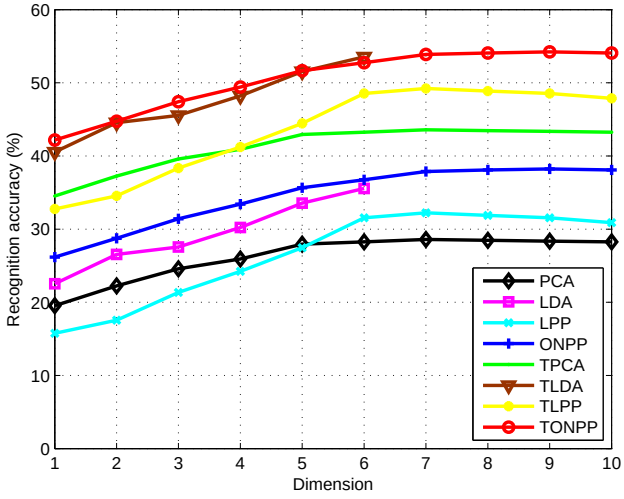


Figure 5.7: Recognition accuracy versus different feature dimensions under the F2C experimental setting.

Table 5.2: Confusion matrix of seven-class expression recognition obtained by PCA under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**30.3%**	22.3%	6.4%	1.4%	20.3%	7.3%	12.0%
DIS	3.6%	**29.8%**	22.4%	1.8%	22.6%	14.3%	5.5%
FEA	13.6%	21.4%	**27.6%**	21.8%	5.2%	7.4%	3.0%
HAP	3.0%	13.6%	21.4%	**28.6%**	20.8%	8.2%	4.4%
SAD	8.6%	16.2%	18.8%	5.6%	**29.6%**	15.2%	6.0%
SUR	3.0%	13.2%	21.8%	18.8%	9.2%	**28.6%**	5.4%
NEU	6.4%	15.3%	12.4%	11.4%	10.2%	19.3%	**25.0%**

samples are independent and identically distributed and this assumption does not hold for cross-dataset facial expression recognition tasks.

The confusion matrices of the seven expressions under the F2C setting were also calculated for PCA, LDA, LPP, ONPP, TPCA, TLDA, TLPP and TONPP, and tabulated in Tables 5.2-5.9, respectively, where ANG, DIS, FEA, HAP, SAD, SUR and NEU represent the anger, disgust, fear, happy, sad, surprise and neutral expressions. We can observe from these results that diagonal elements of the confusion matrices of transfer subspace learning methods are generally better than those of conventional non-transferred subspace learning methods, which further indicates that transfer subspace learning approach can improve the recognition accuracy of subspace learning for cross-dataset facial expression recognition.

5.5 Summary

We have investigated in this chapter the problem of cross-dataset facial expression recognition. To the best of our knowledge, this problem is seldom addressed in

Table 5.3: Confusion matrix of seven-class expression recognition obtained by LDA under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**37.0%**	20.3%	6.4%	1.4%	16.3%	7.3%	11.3%
DIS	3.6%	**36.5%**	20.4%	1.8%	20.6%	12.3%	4.8%
FEA	10.6%	20.4%	**34.3%**	19.8%	5.2%	7.2%	2.5%
HAP	3.0%	11.6%	19.4%	**35.3%**	18.8%	7.7%	4.2%
SAD	6.6%	15.2%	16.8%	5.6%	**36.3%**	14.2%	5.3%
SUR	3.0%	12.2%	19.8%	16.8%	8.2%	**35.3%**	4.7%
NEU	5.4%	13.3%	10.4%	11.4%	8.5%	17.3%	**33.7%**

Table 5.4: Confusion matrix of seven-class expression recognition obtained by LPP under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**34.0%**	20.3%	6.4%	1.4%	19.3%	7.3%	11.3%
DIS	3.5%	**33.5%**	20.4%	1.8%	21.6%	14.2%	5.0%
FEA	12.6%	20.4%	**31.3%**	20.8%	5.0%	6.9%	3.0%
HAP	3.0%	12.6%	20.4%	**32.3%**	19.1%	8.2%	4.4%
SAD	7.6%	15.2%	17.8%	5.3%	**33.3%**	14.8%	6.0%
SUR	3.0%	12.2%	20.8%	17.8%	8.5%	**32.3%**	5.4%
NEU	5.4%	15.3%	11.4%	9.4%	11.2%	18.6%	**28.7%**

Table 5.5: Confusion matrix of seven-class expression recognition obtained by ONPP under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**39.6%**	17.3%	6.4%	1.4%	16.3%	7.2%	11.8%
DIS	3.6%	**39.1%**	18.4%	1.2%	20.6%	12.3%	4.8%
FEA	10.6%	18.4%	**36.9%**	19.2%	5.2%	7.2%	2.5%
HAP	3.0%	11.6%	18.4%	**37.9%**	17.2%	7.7%	4.2%
SAD	6.6%	15.2%	15.8%	5.6%	**38.9%**	13.2%	4.7%
SUR	3.0%	12.2%	17.8%	16.2%	8.2%	**37.9%**	4.7%
NEU	5.4%	13.3%	10.4%	10.8%	8.5%	15.3%	**36.3%**

Table 5.6: Confusion matrix of seven-class expression recognition obtained by TPCA under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**45.3%**	12.3%	6.4%	1.4%	15.3%	7.3%	12.0%
DIS	3.6%	**44.8%**	12.4%	1.8%	17.6%	14.3%	5.5%
FEA	13.6%	11.4%	**42.6%**	16.8%	5.2%	7.4%	3.0%
HAP	3.0%	13.6%	11.4%	**43.6%**	15.8%	8.2%	4.4%
SAD	8.6%	11.2%	13.8%	5.6%	**44.6%**	10.2%	6.0%
SUR	3.0%	13.2%	11.8%	13.8%	9.2%	**43.6%**	5.4%
NEU	6.4%	10.3%	12.4%	10.4%	10.2%	10.3%	**40.0%**

Table 5.7: Confusion matrix of seven-class expression recognition obtained by TLDA under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**55.0%**	10.3%	6.4%	1.4%	10.3%	7.5%	9.1%
DIS	3.6%	**54.5%**	10.4%	1.8%	12.6%	12.3%	4.8%
FEA	10.6%	10.4%	**52.3%**	12.8%	5.2%	6.2%	2.5%
HAP	3.0%	10.6%	10.4%	**53.3%**	10.8%	7.7%	4.2%
SAD	3.6%	10.2%	10.8%	5.6%	**54.3%**	10.2%	5.3%
SUR	3.0%	12.2%	9.8%	8.8%	8.2%	**53.3%**	4.7%
NEU	5.4%	10.3%	6.4%	10.4%	8.5%	7.3%	**51.7%**

Table 5.8: Confusion matrix of seven-class expression recognition obtained by TLPP under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**51.0%**	10.3%	6.4%	1.4%	12.3%	7.3%	11.3%
DIS	3.6%	**50.5%**	10.4%	1.8%	14.6%	14.1%	5.0%
FEA	12.6%	10.4%	**48.3%**	13.8%	5.0%	6.9%	3.0%
HAP	3.0%	12.6%	10.4%	**49.3%**	12.1%	8.2%	4.4%
SAD	7.6%	10.2%	10.8%	5.3%	**50.3%**	9.8%	6.0%
SUR	3.0%	12.2%	10.8%	10.8%	8.5%	**49.3%**	5.4%
NEU	5.4%	1.3%	9.4%	9.4%	10.2%	8.6%	**55.7%**

Table 5.9: Confusion matrix of seven-class expression recognition obtained by TONPP under the F2C setting.

	ANG	DIS	FEA	HAP	SAD	SUR	NEU
ANG	**55.6%**	9.3%	6.4%	1.4%	8.3%	7.2%	11.8%
DIS	3.6%	**55.1%**	11.4%	1.2%	11.6%	12.3%	4.8%
FEA	10.6%	10.4%	**52.9%**	11.2%	5.2%	7.2%	2.5%
HAP	3.0%	11.6%	11.4%	**53.9%**	8.2%	7.7%	4.2%
SAD	6.6%	10.2%	11.8%	5.6%	**54.9%**	8.2%	2.7%
SUR	3.0%	7.2%	11.8%	11.2%	8.2%	**53.9%**	4.7%
NEU	5.4%	11.3%	6.4%	6.8%	8.5%	9.3%	**52.3%**

literature. Since the training and testing samples are not independent and identically distributed in many real facial expression recognition applications, we have proposed a new transfer subspace learning approach to learn a feature space which transfers the knowledge gained from the training set to the target (testing) data to improve the recognition performance under cross-dataset scenarios. Following this idea, we have formulated four new transfer subspace learning methods, i.e., transfer PCA (TPCA), transfer LDA (TLDA), transfer LPP (TLPP), and transfer ONPP (TONPP) for cross-dataset facial expression recognition. Experimental results have demonstrated the efficacy of the proposed methods.

For our future work, we want to explore other facial representation methods such as local binary patterns (LBP) and Gabor features to obtain more robust and discriminative features for transfer learning to further improve the recognition accuracy of cross-dataset facial expression recognition. Moreover, we also plan to implement our proposed approach for practical human robot interaction applications to further show its effectiveness.

Chapter 6

Dorothy Robotubby Evaluation in Real Pilot Studies

6.1 Introduction

With the rapid development of social robots and increasing demands from robotic users, human-robot interaction has been a hot topic and received a growing interest in social robotic area over the past five years. As Goodrich and Schultz [90] defined:"Human-Robot Interaction (HRI) is a field of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans". In this chapter, we evaluate our developed robotic nanny named Dorothy Robotubby based on HRI.

An accurate evaluation not only reflects the developed robot's performance such as the usability, robustness, timeliness, and automaticity, but also provides the feedback information from the users to help robotic designers to develop acceptable and satisfactory robotic systems. That is because the users are true and final operators of the robots. A feasible and popular solution to achieve effective

evaluation results is pilot studies or field trials which test the robot with its target subjects in lab environments or directly in real application environments. This method has been used in many robot evaluation processes. For example, Keepon was conducted a pilot study for its rhythmic interaction with children in the lab [47]; Olivia, a social robot that can inform and entertain visitors, was tested by interacting with 120 visitors in a two-day annual exhibition *TechFest* [91]; Iromec [50] was placed at the primary school G.Pascoli and S.Martini in Siena to evaluate the design aspects of the robot such as usability and acceptability; and Paro was introduced to a care house to study its sociopsychological and physiological influences on the elderly [45].

Dorothy Robotubby was introduced in Chapter 3 to play with and take care of a child in case his/her parent or caregiver is absent. There are two main user interfaces in our Dorothy Robotubby system: local control-based and remote control-based. Local control-based interface is developed for a child to control the robot directly to execute some tasks such as telling a story, playing music and games, chatting, and video calling. Remote control-based interface is designed for parents to control the robot remotely to execute several commands like demonstrating facial expressions and gestures.

In this chapter, the used pilot studies focus mainly on two aspects: 1) to evaluate whether the children like the appearance and functions of Dorothy Robotubby, and 2) to collect the parent's opinions on the remote user interface designs. Have analyzed the pilot study results, in addition to the general evaluation of Robotubby's performance, the feedback information from children and parents can help us to reposition the developed robot such as the user's age range and the robot's

Table 6.1: Personal information of the children involved in the survey.

Child No.	1	2	3	4	5	6	7
Age	4	5	5	10	10	12	13
Gender	Female	Female	Female	Male	Female	Male	Male
Q1	Yes	Yes	Yes	No	Yes	No	Yes
Q2	No	No	Yes	No	Yes	Yes	Yes
Note	Q1: Are you interested in a robot? Q2: Are you familiar with or see a robot before?						

application areas. Moreover, these useful information can provide significant reference to improve the current functions and to design new functions for our robot. Since the functions of video call and remote control in remote user interface only occupy a small part in the whole robot system, we mainly evaluate the interaction between the child and the robot.

This chapter is organized as follows. Section 6.2 introduces experimental settings and procedures. Section 6.3 describes evaluation methods. Section 6.4 discusses experimental results, and Section 6.5 concludes this chapter.

6.2 Experimental Settings and Procedures

We conducted pilot studies in the Control and Mechatronics Lab at National University of Singapore. Seven children and five parents who are friends of our group were invited to help us to test our robot. There are 4 females and 3 males aged from 4 to 13 years old. Table 6.1 lists some personal information of these children. Among these children, 5 are interested in a robot and 4 are familiar with or have seen a robot before. The tests were organized in individual session except for two children who are the youngest with 4 and 5 years old, respectively.

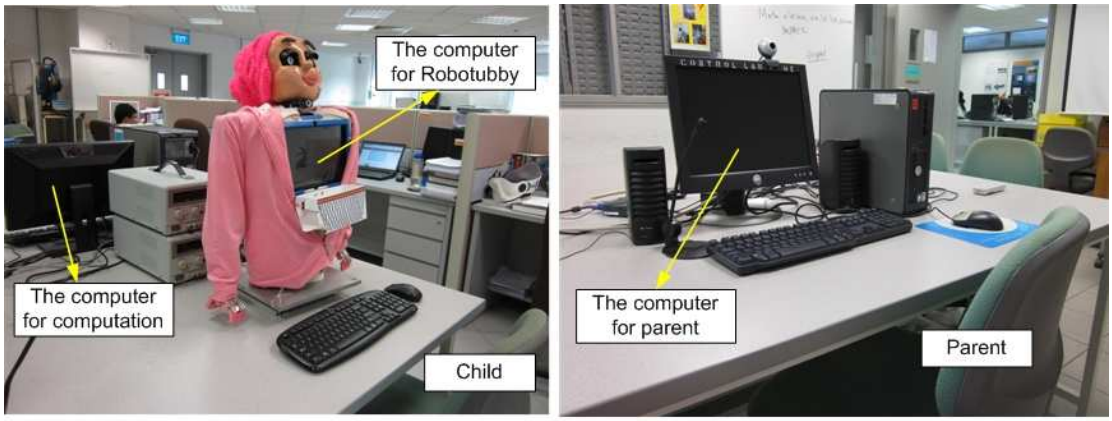


Figure 6.1: Two testing rooms of pilot study where (a) is testing room for the child and (b) is testing room for the parent.

Considering that the children with 4 or 5 years old may feel uncomfortable to a new environment, we allowed the parent to attend with their children at the beginning of testing. Each child usually requires 25-30 minutes to complete the test. Since it is difficult to ensure that all the involved children and parents are available at the same time, we arranged their testings at three different sessions, where 4, 2 and 1 children test our system in different sessions, respectively.

There are three computers used in our robot system. The computer for robot and the computer for computation were placed in one room, and the computer for parent was placed in another room. The distance between these two rooms is far enough to ensure that the child and parent will not see and hear each other when they test the video call and remote control functions. Figure 6.1 shows the two testing rooms. A brief introduction on the experiments was presented to the child and parent at the beginning of the test. During the whole test, a human assistant also participated to answer questions and help solve problems from the child and

parent. For younger children with 4 or 5 years old, an assistant helped them to operate the robot until they could do it by themselves. While for the children older than 10 years old, the assistant normally encouraged them to explore the robot by themselves first. If required, the human assistant would supervise the children's activity.

There were two testing parts in the whole experiment which were designed to follow a certain sequence. Firstly, the child was requested to test the robot's functions including story telling, chatting, music playing, game playing, face tracking, and emotion recognition. Secondly, the parent was asked to test the remote user interface's functions including video call and remote control with the child. During the whole test, the interaction activities between the child and the robot were observed and recorded. After testing, participants were asked to complete a questionnaire.

6.3 Evaluation Methods

To effectively evaluate and analyze the performance of Robotubby and the interaction between the child and the robot, the questionnaires focusing on children and parents were prepared, respectively. The questionnaire for child was designed to evaluate the robot's functions and appearance, and the child's feelings during the interaction. The questionnaire for parent was employed to investigate the parent's feelings about the remote user interface's design. The questions on both questionnaires were based on a 5-point Likert scale and some suggestions were requested if possible. Tables 6.2 and 6.3 enumerate the questions used in the

Table 6.2: The questions used in the questionnaire for the child.

Question1	Robotubby system includes the functions of story telling, chatting, music playing, game playing, face tracking, video call, emotion recognition, and remote control, please score every one according to their usability.
Answer	very unsatisfied; unsatisfied; normal; satisfied; very satisfied
Question2	For the above functions, which one or ones do you like?
Answer	Storytelling; Chatting; Music playing; Game playing; Face tracking; Video call; Emotion recognition; Remote control
Question3	About our Robotubby, how do you find its appearance?
Answer	very scary; scary; normal; appealing; very appealing
Question4	What is your feeling to Robotubby after interaction with it?
Answer	very boring; boring; normal; interesting; very interesting
Question5	Do you think Robotubby can be your friend?
Answer	totally cannot; cannot; maybe; can; totally can
Question6	Do you think that Robotubby can appropriately recognize your emotional states and feelings when it tells a story?
Answer	totally cannot; cannot; maybe; can; totally can

Table 6.3: The questions used in the questionnaire for the parent.

Question1	Please score remote user interface from the factors of appearance, operability, and functions.
Answer	very unsatisfied; unsatisfied; normal; satisfied; very satisfied

questionnaires for the child and parent, respectively.

Summarizing the answers to each question in the questionnaires can reveal the direct attitudes to the robot from both children and their parents. For example, whether they like or dislike the developed robot. Analyzing the suggestions from the children and parents can help us to understand their expectations to the robot and thus to narrow the gap between our robot and their satisfactory robots.

Since the results from questionnaires are normally subjective, in addition to using questionnaires, the activities of children with the robot were recorded by a video

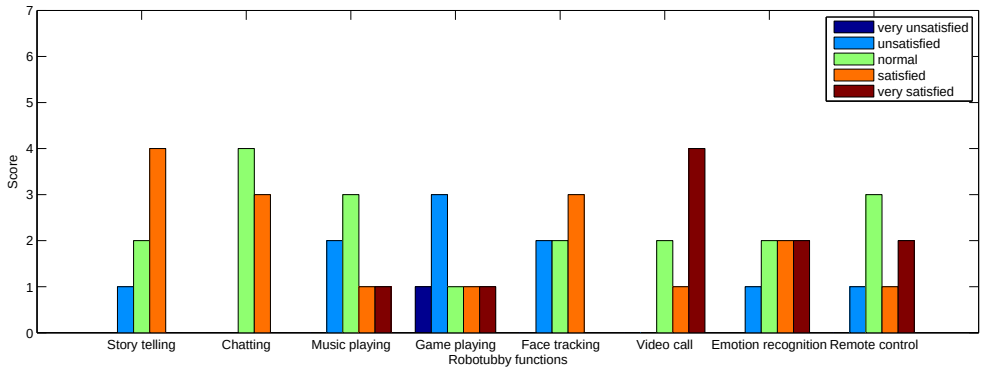


Figure 6.2: The statistical result of Question 1 in Table 6.2.

camera during the testing to increase objectivity of evaluation. By analyzing body gestures, facial expressions, and verbal behaviors of the children in videos, we can obtain more detailed information on the children's feelings to Robotubby. Such behavior analysis method has been widely used to study human-robot interaction and also commonly applied in psychology to acquire knowledge in human social interactions [91].

6.4 Results and Discussion

6.4.1 Results from Questionnaire Analysis

As listed in Table 6.2, the questionnaire for the child includes 6 questions that focus mainly on the evaluation to the robot's functions and appearance, and the feelings of the child to the robot. Figures 6.2-6.7 illustrate the statistical result of each question in the questionnaire based on the children's assessments. Here the score values indicate the number of children who vote for this category and hence the maximal value should be the total number of children (7) involved in the test.

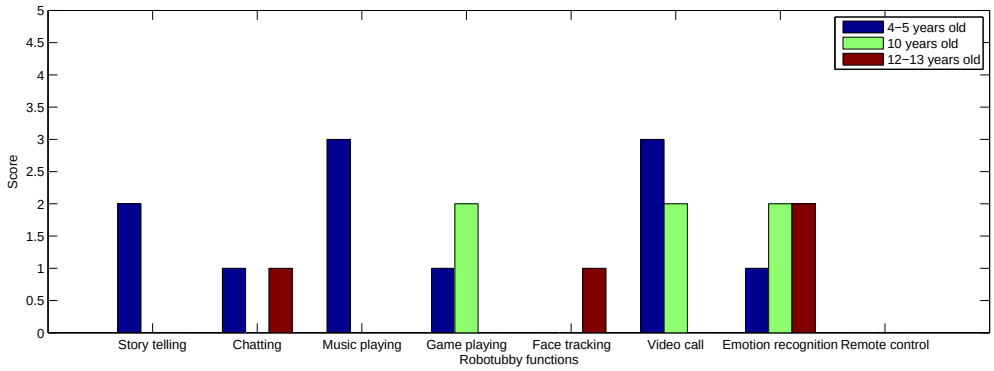


Figure 6.3: The statistical result of Question 2 in Table 6.2.

Figure 6.2 shows the children's responses to Question 1 in Table 6.2. The aim to design this question is to evaluate each function of Robotubby according to usability. We used 5 different colors to represent the degree of satisfaction. Dark blue, light blue, green, orange, and brown are employed to denote "very unsatisfied", "unsatisfied", "normal", "satisfied", and "very satisfied", respectively. From the figure, it can be seen that for the functions of chatting and video call, there is no children to give negative assessments like "very unsatisfied" and "unsatisfied"; for the functions of story telling, emotion recognition, and remote control, one child is unsatisfied; for the functions of music playing and face tracking, two children are unsatisfied; and for game playing, four children are unsatisfied, and one of them is very unsatisfied. Generally speaking, most children feel normal or satisfied to the developed functions except for game playing. The reason why game playing is unsatisfied may be that there is only one game in this function. To improve this function, one child suggested to add more game types, and another child advised to make the game playing faster.

Figure 6.3 shows the children's responses to Question 2 in Table 6.2. The colors

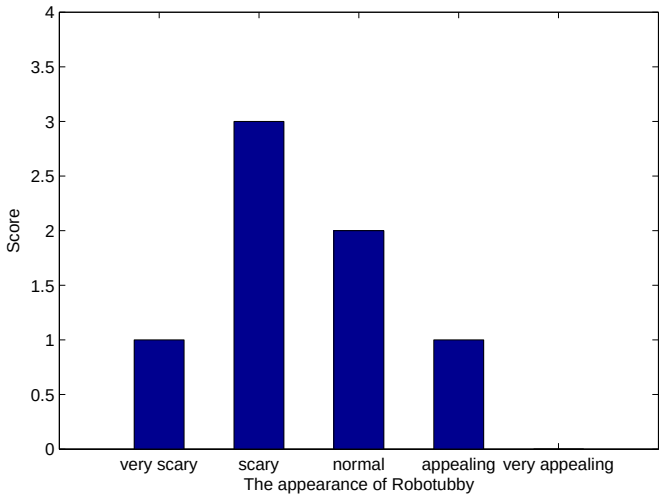


Figure 6.4: The statistical result of Question 3 in Table 6.2.

of dark blue, green, and brown represent the age groups of the children. The results show that the children involved in the survey show interests in different robot functions versus their different ages. In particular, the children from 4 to 5 years old usually like the functions of story telling, music playing, and video call; the children with 10 years old normally like the functions of game playing, video call, and emotion recognition; the children from 12 to 13 years old like the functions of chatting, face tracking, and emotion recognition; and no children like the remote control function. This may be because the selected stories and music are more suitable for younger children; the designed game is comparatively easy for the children older than 10 years; younger children are more happy to talk with their parents; older children usually like the function with higher technologies; and less children like to be interfered when interacting with the robot.

For the appearance of the robot, the evaluation result is shown in Figure 6.4. We can find that almost half of the children think it is scary and the rest think it is

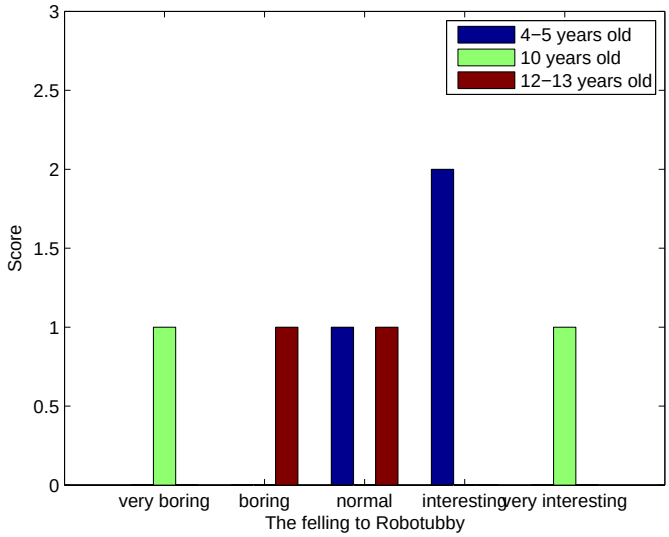


Figure 6.5: The statistical result of Question 4 in Table 6.2.

normal or appealing. Two remarkable appearance features of Robotubby are the highly mobile face and the touch screen mounted on its belly, which are the main reasons why children think our robot is appealing. While some children think the robot is scary, a possible explanation is that the colors of the robot's skin and eyeballs are abnormal compared with real humans. One child who dislikes the robot's appearance gave his own reason:"the eyes are too big and never blink, therefore, you will feel like someone is watching you." Correspondingly, this child suggests us to add a blinking function to the robot's eyes. Two more children think it would be better if the robot has hair and legs.

Figures 6.5 and 6.6 illustrate the opinions of the children after interacting with Robotubby. From these two figures, we can find that five children think the robot is normal or interesting after interacting with it and four children think the robot maybe or can be their friend. It can also be found that the children who think

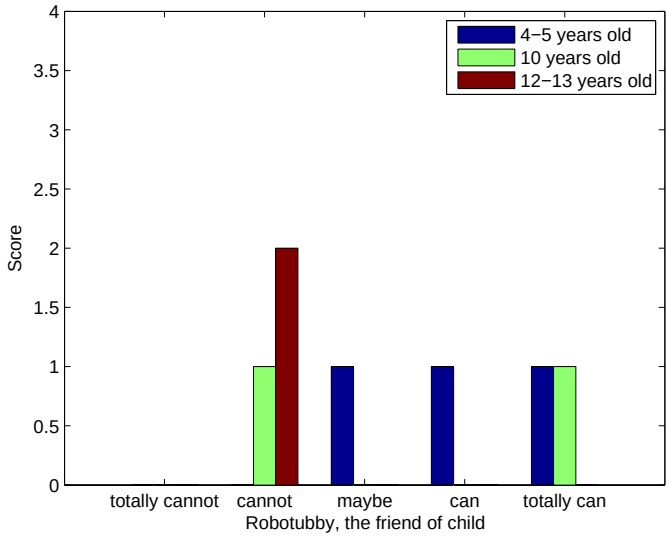


Figure 6.6: The statistical result of Question 5 in Table 6.2.

Robotubby is boring and cannot be their friends are generally older than 10 years old. Compared with younger children, they may have higher requirements for the developed robot. Moreover, these older children normally have their own interest. For example, two children gave the reasons why they think the robot is boring. One child said:"It is not a car!" And the other child said:"I am not so interested in music, stories, and basketball." This may affect the attitude of a child to the robot.

Generally speaking, most of the existing functions of Robotubby are too simple for the older children and it is difficult to maintain their interest in the robot. On the other hand, the younger children are more interested in Robotubby. Therefore, our developed robot is more suitable for the younger children usually from 5 to 10 years old.

Since emotion recognition plays important roles on social robotics for a child and

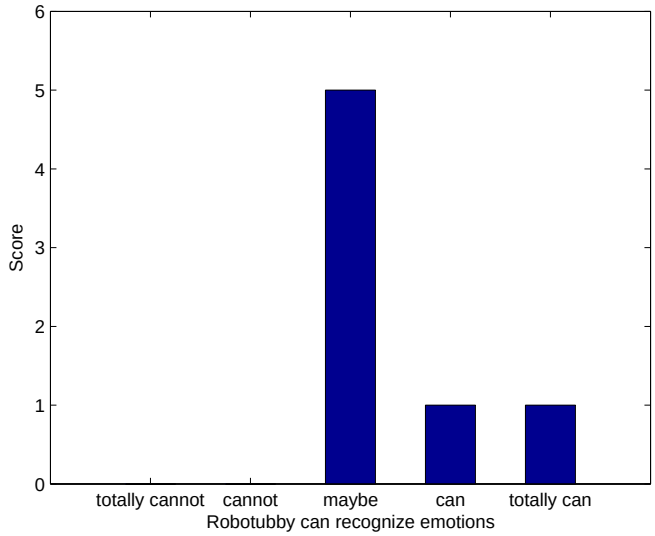


Figure 6.7: The statistical result of Question 6 in Table 6.2.

it is also an important part in the whole thesis, we designed a question (Question 6 in Table 6.2) to individually evaluate the effect of emotion recognition in Robotubby, as shown in Figure 6.7. We can see from the figure that all the children thought that Robotubby maybe or can appropriately recognize their emotional states and feelings when it tells a story. One of the children said:“ It knows that I was somewhat depressed.” Although emotion recognition is only applied for storytelling function of the robot with current version, it still can activate the children's interest to interact with the robot. That is because once the robot is not perceived as a mere machine due to its emotion recognition function in storytelling, the children may easily keep it in their minds during the whole procedure of interaction with the robot. Emotion recognition function makes the behavior of the robot more believable and acceptable.

Compared with the survey on the children, we only prepared a simple question for

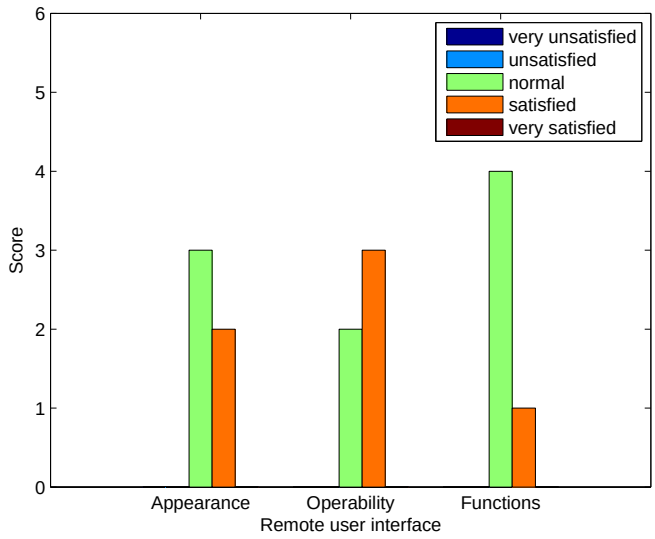


Figure 6.8: The statistical result of Question 1 in Table 6.3.

the parents to evaluate the appearance, operability, and functions of the remote user interface. Figure 6.8 shows the result of Question 1 in Table 6.3. As a general observation from the figure, the parents thought the designed remote user interface was normal or satisfactory. Considering the degree of satisfaction of each evaluation item, three out of five parents thought operability of the remote user interface is satisfactory which has the highest degree of satisfaction. The followings are appearance and function items.

While there is no negative assessment with regard to the remote user interface's design, the parents offered us several suggestions. These suggestions not only reflect the parents' expectations to the design of remote user interface, but also help us to further improve it.

With regard to appearance design, one parent suggested that the interface should occupy the entire screen and some components should be highlighted such as the

information showing robot status. The advice about operability design came from another parent who thought there are too many buttons on the interface and the simpler the better. Most of the rest suggestions focused on the interface's function design. In summary, three persons recommended to add another camera to show the whole scene of the child and the robot together such that more surrounding information and visual feedback could be provided to the remote side. One parent proposed that the sound system in video call should be improved.

6.4.2 Results from Behavior Analysis

The interaction between the child and the robot was recorded by a camera. After the entire testing on 7 children, we replayed the recorded videotapes and annotated the participants' behaviors. To increase the reliability of the obtained results, we used two coders for annotation. The behaviors can be mainly classified into two categories according to the degree of participation: high-interactive and low-interactive. High-interactive behaviors include gaze, smile, touching, and speech communication. Low-interactive behaviors consist of looking at the left and right without focusing attention, quietly sitting with depressed expression, and operating with no expression. We analyze these behaviors in this subsection.

Gaze behavior: During the interaction, the most frequent behavior is the gaze behavior. Children's gaze behavior can be described as gazing predominantly at the robot, gazing predominantly at the screen, and mixed gazing at both screen and robot. The first type of gaze behavior normally appeared when the robot demonstrated different facial expressions and gestures. The second type usually occurred when the child operated the robot by clicking the buttons on the screen

(a) (b)

Figure 6.9: Two examples of the children's gaze behavior.

or the content on the screen changed. For the last type, it generally occurred when the robot's facial expressions or gestures and the screen's content simultaneously changed such as executing story telling and music playing functions. Figure 6.9 shows two examples of the children's gaze behavior.

Contrary to the gaze behavior, three children occasionally behaved in the manner of looking at the left and right without focusing attention during the interaction. The recordings revealed that under such condition, the children's attentions were distracted from the robot to the other external factors such as the sound of voices and the actions from other persons. It also implied that the currently executed function of the robot cannot attract and maintain the children's attention.

Smile behavior: For humans, smile is an expression denoting pleasure, joy, happiness, or amusement. During the interaction, smile is the child's response to the robot's behavior. All the children expressed this behavior during their interactions. The difference is their different duration in smiling. Roughly speaking, child 1 (C1), C2, and C5 smiled more than other children. C4 and C6 seldom smiled during the interaction except when they operated chatting and video call

(a) (b)

Figure 6.10: Two examples of the children's smile behavior.

functions of the robot. Figure 6.10 illustrates two examples of the children's smile behavior.

In addition, we examined these children's questionnaires again and found a significant correlation between smile times and the child's feelings to the robot. Generally, the children who express more smile behavior thought the robot is interesting and can be their friend. In contrast, the children who seldom smile such as C4 and C6 thought the robot is boring and cannot be their friend.

Touching behavior: Besides the above two behaviors, touching is another action used by the children during the interaction. The children's touching behaviors include touching the robot's hands and touching the robot's face. Through observing the recorded videos, we found that not all the children expressed this behavior. Among the children, C2 and C3 touched the robot's hands, and C3 and C7 touched the robot's face such as its skin, eyebrows, and mouth. By combining facial expression and gesture analysis for these children, we think that C2 and C3 were close to the robot and expected to interact with it when they touched the robot's hands, and C3 and C7 seemed curious to the robot's facial components

(a) (b)

Figure 6.11: Two examples of the children's touching behavior.

when they touched the robot's face. While there is not a direct relationship between touching behavior and the feelings of children to the robot, this behavior could show that these children focused their attention on the robot at the moment of expressing touching behavior. Figure 6.11 shows two examples of the children's touching behavior.

Speech behavior: Similar to touching behavior, speech is an occasional behavior existing in interaction. Usually, the children asked for help from the assistant via speech when they did not know how to explore the functions of the robot or got into difficulty during interacting with the robot. In addition, some children asked questions about the robot and provided advices to the assistant through speech during the interaction. For example, C1 asked the assistant:" Why dose the robot have no legs?" C5 said:" The game is so easy and I think it would be better to throw the ball at a longer distance like at a 2-meter distance." C5 also said: "The younger children may like the music in the robot. It should allow the users to choose their own songs because different people may like different songs." Compared with these questions and suggestions, the encouraging thing

is that sometimes the children used greeting and polite languages to the robot. For instance, C2 and C5 said “Hello, robot” when they saw the robot at the first time. After the robot accomplished a task or it passed the ball to the child, C5 said “ Thank you” and “ Thank you very much” to the robot. When the children express such speech behavior, they may enjoy the interaction with the robot at this moment.

Besides gaze, smile, touching, and speech behaviors, the children also expressed some other low-interactive behaviors such as quietly sitting with depressed expression and operating with no expression. These behaviors were normally expressed by C4 and C6. These two children are not interested in the robot. It implies that similar to smile behavior, these two behaviors may show some significant correlations with the child’s feelings to the robot. If the children have no interest in the robot, they will easily and frequently express these two behaviors.

When improving the current functions and designing new functions of the robot, it would be better to consider how to activate the children to behave more interactively. Due to the relationship between the above mentioned behaviors and the degree of interaction with the robot, a reasonable solution is to develop the functions that let the children frequently express the highly-interactive behaviors such as gaze, smile, touching, and speech. For example, making the functions more various and changeable may easily catch the child’s eyes like adding pictures to storytelling; giving proper tactile feedbacks by the robot may make the child be willing to touch the robot; and adding speech recognition function may increase the child’s interest to talk with the robot.

6.4.3 Results from Case Study

To comprehensively study the interaction between the child and the robot, we analyzed three cases including four children. The reason to choose these four children is that they expressed various behaviors during the interaction.

Case 1: The first case is a four-year-old girl C1 and a five-year-old girl C2. They are from the same family. Since they are young, we arrange their mother to interact with the robot together. When these two girls came to our lab and saw the robot at the first time, they were excited and instantly sat before the robot. After the assistant told them the robot can tell a story, C1 happily said:“ Oh, great! It can tell a story for us.”, as illustrated in Figure 6.12(a).

However, when the robot began to tell a story with different gestures, the children were apparently scared by the robot's sudden motion. They ran away from the robot and hid behind a chair. To release the children's fear to the robot, we stopped the story and played a music to them. While the children still did not dare to approach the robot, they straight gazed the robot as if they were attracted by the robot's song and dance. Slowly, C2 did not hide behind the chair any more and just stood away to watch the robot. This scene is shown in Figure 6.12(b).

To ensure the children interact with the robot successfully, we advised the children to play a game with the robot. Their mother also told them the robot will not hurt them and encouraged them to play with the robot. Then they sat before the robot with their mother. After several demonstrations from the assistant, the children gradually played the game by themselves which is shown in Figure 6.12(c).

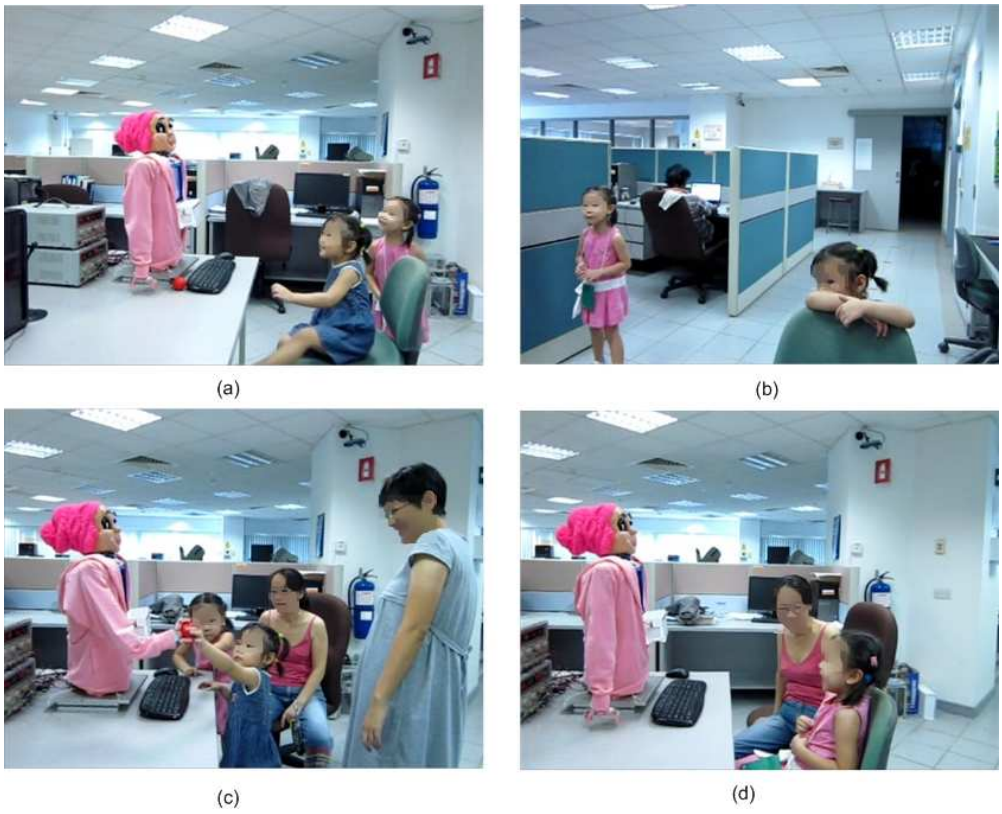


Figure 6.12: Several pictures for Case 1.

When testing the functions of chatting and video call, the children totally accepted the robot and were not scared any more. They began to touch the robot's hands and follow its gestures. Frequent smiles indicated that they enjoyed to interact with the robot. Figure 6.12(d) shows this scene.

The first case described the procedure of the children's attitude changes to the robot: from scare and acceptance to enjoyment. By looking up these two children's questionnaires and observing their videotapes, we found that they liked storytelling function of the robot. What scared the children seemed the sudden motion of the robot that may be beyond their imagine. In addition, young age

and unfamiliarity to the robot are also possible reasons. Motivated by the fact that with gradual familiarity to the robot, these two children finally enjoyed the interaction, we could add a self-introduction function into the robot. The self-introduction of the robot could be scheduled before the interaction by means of several short videos which could be pre-stored in the computer of the robot. That may make the children familiar with the robot in advance and reduce their fear when interacting with the robot [44].

Case 2: The second case is a 10-year-old girl C5. From the beginning to the end of the testing, she behaved actively and interested in Robotubby. When this girl came to our lab and saw the robot at the first time, she said "Hello" to the robot. After the assistant introduced the basic functions of the robot to her, the child requested to play Mini game first as she liked to play games very much. At the beginning of the game, the girl totally attracted by the robot, especially finding that the robot can automatically pick up the ball from the basket and then pass the ball to her. She expressed her affection by clapping hands and smile. Figures 6.13 (a) and (b) show these two behaviors of C5, respectively. She also used the words " This robot is cute!" to commend Robotubby's actions and said "Thank you!" to the robot as if the robot is not just a machine. With several times' repeats of this game, the child was not so interested like the beginning any more. She though the game is easy for the children at her age (10 years old) and it would be better to increase its difficulty such as throwing the ball at a longer distance.

Then the child C5 tried storytelling and music playing functions of the robot. When telling a story by the robot, C5 carefully listened to the story. Her gaze changed between the story content of the screen and the demonstrated gestures

Figure 6.13: Two examples of C5's behavior for Case 2 where (a) is clapping hands and (b) is smile.

of the robot. Sometimes, she touched the robot's hands or face and smiled to the robot. She said nothing about this function. By observing and analyzing her behavior, while the child acted interactively to a certain extent during storytelling, she did not show much interest in this function since it is difficult to detect excited expressions from her behavior. During playing music video, the child smiled several times and provided us some suggestions. She thought younger children may like the selected music and it would be better to let the user choose their favorite songs.

Next, the girl tested the functions of chatting and video call with her mother. For chatting function, she asked the robot a question of "3+7=?" by means of speech and typing texts, respectively. After finding the robot cannot correctly answer, she smiled and then said "I beg your pardon" with a little disappointed. At this time, it seemed that the girl did not treat Robotubby as a machine and she expected that the robot could understand her words and help her to solve the problems like a real human. With regard to video call, she expressed much interest

again. She enjoyed to talk with her mother through the robot. In addition, she followed the robot's actions from remote control and guessed what the action is.

The second case is a representative example that the child shows great interest in the robot and is well attracted by the robot. The result can be obtained by observing and analyzing the child's behavior. During the whole procedure of interaction with the robot, the child C5 frequently expressed highly-interactive behaviors that have been mentioned in the above subsection such as gaze, smile, touching, and speech. In addition, she seldom behaved like looking at the left and right without focusing attention, quietly sitting with depressed expression, and operating with no expression. The similar result also can be obtained by looking up C5's questionnaire. The questionnaire showed that this child felt the robot is very interesting after interacting with it and also thought that the robot totally can be her friend. The answers from the questionnaire are consistent with the expressed behaviors by the child.

Through studying the second case and summarizing the suggestions given by C5, we could find that even if the child shows great interest in the robot and is well attracted by the robot, we still need to improve or develop more various and changeable functions of the robot. This is because we just arrange the children to interact with the robot in a short time. With the passage of time, the children may gradually lose their interest if the functions of the robot are changeless, especially for the older children. To maintain the children's interest, the functions suitable for long-term interaction should be considered. In addition, we found that if the robot behaves intelligently such as automatically picking up and passing the ball and chatting with the child, the child does not easily treat the robot as a mere

machine. For example, C5 communicated with the robot with the speech when the robot behaved intelligently. Hence, more intelligent functions of the robot should be taken into account.

Case 3: The third case is a 13-year-old boy C7. Considering the current functions of the robot and the age of the child C7, it seems that the current functions are relatively simple and easy for the child with 13 years old. That may be the reason why C7 did not demonstrate great interest in the robot. This is also verified by observing the videotape and analyzing the questionnaire. We can find from the videotape that during the interaction with the robot, the child only expressed highly-interactive behaviors like gaze, smile, and touching several times. He frequently looked at the left and right without focusing attention. According to the answers from the questionnaire, it can be seen that the child C7 thought the robot is normal after interaction and it cannot be his friend.

While the child did not behave highly-interactively during interacting with the robot, we still selected him as a study case. This is because after his own testing with the robot, the child also participated other children's testing. For example, when other children interacted with the robot, he liked to be around to see them and touched the robot's face sometimes. Moreover, since C7 is the first child to test the robot and it is easier for him to operate the robot at his age, he liked to help other children if they met some difficulties when operating the robot. We did not arrange this testing part for him, and this activity by him is totally spontaneous. This is an interesting behavior and different from other children. For other children, after they finished their own testing parts, some of them would leave the testing location, and some of them would do their own thing without

(a) (b)

Figure 6.14: Two scene examples of C7.

drawing attention to the robot again even if they still were in the testing location. Figures 6.14 (a) and (b) show two scene examples of C7.

The third case C7 demonstrated a special behavior: participation into other children's testing where the activity of helping other children looks like that from the assistant to a certain extent. In addition, compared with other children, the child C7 seemed to like to touch the robot such as its hands, skin, and facial components. The reason why C7 expressed these behaviors may be that he would like to understand more and deeper knowledge about the robot by exploring the robot by himself and watching the interactions of other children. While we do not know the real reason why C7 behaved like that, the child's behavior could prove that he is interested in something related to the robot even if that is not about the developed functions of the robot. It inspires us that Robotubby may be an intermedia between the child and other persons. For instance, the child and his/her friends could play some games together that are developed on the robot, and the teacher could take Robotubby as a tool to help the children learn knowledge.

6.4.4 Discussion

To better evaluate our developed robot, we employed questionnaires and videotapes in our study. The obtained results are from questionnaire analysis, behavior analysis, and case study. Having summarized and analyzed these results, we can find there is a consistency between them. Specifically, the children who are more interested in this robot generally behave high-interactively and the children who think the robot is boring usually show low-interactive behaviors.

As for the children's attitude to the robot and behaviors expressed in the interaction, many factors could influence them such as prior experience to the robot and the preferences of males and females. The children C1 and C2 are examples to show the influence of prior experience to the robot. As listed in Table 6.1, these two children are not familiar with or have not seen a robot before. When they saw the robot at the first time, they were very happy. But when the robot moved, they felt fear. After interacting with the robot for a while, they felt happy again. While for the other children who have interacted with the robot or something similar, when the robot began to move, no one felt fear. Moreover, they can operate the robot well after the helper's introduction.

By analyzing the questionnaires and behaviors of the children, we have found that the attitude and feelings to the robot from males and females are different. Compared with male children, female children gave better evaluation to our robot through the questionnaires. Moreover, through observing the behaviors of the children during the interaction with the robot, it can be seen that female children behaved more actively. The reason may be that female children normally like to

play dolls and male children usually like cars or ball games, while the appearance of our robot is more similar with dolls.

6.5 Summary

In order to improve the current functions and develop new functions of the robot, we have designed a pilot study in this chapter from two main aspects: to evaluate whether the children like the appearance and functions of Dorothy Robotubby and to collect the parents' opinions on the remote user interface designs. In the pilot study, 7 children aged from 4 to 13 years old and 5 parents were invited to our lab to attend this survey. After testing, questionnaires and videotapes were employed to analyze the performance of Robotubby and the interaction between the child and the robot. Results from questionnaire analysis, behavior analysis, and case study have shown that while there is some room to improve our robotic nanny, most children and parents express great interest in our robot and provide comparatively positive evaluation. More important, several valuable and helpful suggestions have been summarized and obtained from the result analysis phase. That could make our robot more fascinating and to be used for more applications.

For future work, we are interested to improve the appearance, functions, and user interfaces of the currently built robot system according to the children's and parents' feedback, and improve the system by designing more effective functions. For instance, a Kinect camera can be used to enable Robotubby to copy and follow the child's and parent's certain gestures. A birds-eye-view camera can be utilized such that the parent could see the whole picture of the interaction between the

child and the robot. In addition, the application for the autistic children with Robotubby will be explored and the functions aiming to the therapy for them will be designed.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we have introduced our designed robotic nanny called Dorothy Robotubby which aims to play with and take care of a child in case his/her parent or caregiver is absent. Since emotion recognition can make important contributions towards achieving a believable and acceptable robot and has become a necessary and significant function in many social robotics for a child, we have also studied facial expression-based emotion recognition and addressed two problems which are important to drive facial expression recognition into real-world applications: misalignment-robust facial expression recognition and cross-dataset facial expression recognition. Lastly, we have evaluated our robot Dorothy Robotubby in a real pilot study. The followings detail the key contributions.

We first developed a robotic nanny named Dorothy Robotubby with a caricatured appearance, consisting of a head, a neck, a body, two arms, two hands, and a touch

screen in its belly. There were two main user interfaces in the designed robotic system: local control-based and remote control-based. Local control-based interface was developed for a child to control the robot directly to execute some tasks such as telling a story, playing music and games, chatting, and video calling. Remote control-based interface was designed for parents to control the robot remotely to execute several commands such as demonstrating facial expressions and gestures. By operating these two interfaces, our robot can not only interact with a child, but also build a connection between a child and his/her parent. In addition, due to the independent development, the built robot could be a robotic platform that is easy to add new functions and explore new applications for the robot.

Second, we proposed a new misalignment-robust subspace analysis approach for facial expression recognition. We first empirically showed that spatial misalignments indeed affect the recognition accuracy of conventional subspace learning-based facial expression recognition methods. To make better use of the different interclass samples in learning the feature subspace, we proposed a biased subspace analysis method by imposing large penalties on interclass samples with small differences and small penalties on those samples with large differences simultaneously such that more discriminative features can be extracted for recognition. Moreover, we learned a robust feature subspace by using the IMage Euclidean Distance (IMED) rather than the widely used Euclidean distance such that the subspace sought is more discriminative and robust to spatial misalignments. Experimental results on two widely used face databases have demonstrated the efficacy of the proposed method.

Then, we investigated the problem of cross-dataset facial expression recognition.

Since the training and testing samples are not independent and identically distributed in many real facial expression recognition applications, we proposed a new transfer subspace learning approach to learn a feature space which transfers the knowledge gained from the training set to the target (testing) data to improve the recognition performance under cross-dataset scenarios. Following this idea, we formulated four new transfer subspace learning methods, i.e., transfer PCA (TPCA), transfer LDA (TLDA), transfer LPP (TLPP), and transfer ONPP (TONPP) for cross-dataset facial expression recognition. Experimental results have demonstrated the efficacy of the proposed methods. Since facial images with misalignment and cross-dataset problems are common in real-world applications, the proposed methods can serve as study reference to drive facial expression recognition into real-world applications.

Lastly, we designed a pilot study to evaluate whether the children like the appearance and functions of Dorothy Robotubby and collect the parents' opinions on the remote user interface design. In the pilot study, we invited 7 children and 5 parents to our lab to attend this survey. After testing, we employed questionnaires and videotapes to analyze the performance of Robotubby and the interaction between the child and the robot. Results from questionnaire analysis, behavior analysis, and case studies have shown that while there is some room to improve our robotic nanny, most children and parents express great interest in our robot and provide comparatively positive evaluation. More important, several valuable and helpful suggestions have been obtained from the result analysis phase. That could make our robot more fascinating in more applications in the future.

7.2 Future work

In this section, we present some research directions which can be explored in the future.

For misalignment-robust facial expression recognition, we will further extend the proposed misalignment-robust subspace analysis approach to other supervised manifold learning methods to further explore the nonlinear manifold structure of facial expression data. Moreover, how to design a better penalty function to further improve the recognition performance remains another interesting direction of future work. We are also going to collect more facial expression images under uncontrolled environments to examine the robustness of our proposed method in real-world applications. In this study, we only assume there is spatial misalignment in facial images, however, this assumption may not hold because there could be some other variations in facial expression images such as varying illumination, poses, and occlusions, even for the same person. Hence, how to simultaneously deal with the spatial misalignment as well as other variations for robust facial expression recognition remains to be addressed in the future.

For cross-dataset facial expression recognition, we want to explore other facial representation methods such as local binary patterns (LBP) and Gabor features to obtain more robust and discriminative features for transfer learning to further improve the recognition accuracy of cross-dataset facial expression recognition. Moreover, we also plan to implement our proposed approach for practical human robot interaction applications to further show its effectiveness.

For our robot Dorothy Robotubby, we are interested to improve the appearance,

functions, and user interfaces of the currently built robot system according to the children's and parents' feedback, and improve the system by designing more effective functions. For instance, a Kinect camera can be used to enable Robotubby to follow the child's and parent's certain gestures. A birds-eye-view camera can also be utilized such that the parent could see the whole picture of the interaction between the child and the robot. In addition, the application for the autistic children with Robotubby is another interesting direction to be explored in the near future.

Bibliography

[1] C.L. Breazeal. *Designing sociable robots*. The MIT Press, 2004.

[2] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3):143–166, 2003.

[3] C. Bartneck and J. Forlizzi. A design-centred framework for social human-robot interaction. In *International Workshop on Robot and Human Interactive Communication*, pages 591–594, 2004.

[4] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer. Understanding social robots. In *International Conferences on Advances in Computer-Human Interactions*, pages 169–174, 2009.

[5] Social robot. `http://en.wikipedia.org/wiki/Socialrobot`, 2011. Last accessed on Nov 5th 2011.

[6] N. Sharkey and A. Sharkey. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies*, 11(2):161–190, 2010.

[7] J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell. The clinical use of robots for individuals with autism spectrum disorders: a critical review. *Research in Autism Spectrum Disorders*, 6:249–262, 2012.

[8] E.L. Broek. Robot nannies: Future or fiction? *Interaction Studies*, 11(2):274–282, 2010.

[9] S. Turkle, C. Breazeal, O. Dasté, and B. Scassellati. Encounters with kismet and cog: Children respond to relational artifacts. *Digital Media: Transformations in Human Communication*, pages 1–20, 2006.

[10] E. Shin, S.S. Kwak, and M.S. Kim. A study on the elements of body feature based on the classification of social robots. In *International Symposium on Robot and Human Interactive Communication*, pages 514–519, 2008.

[11] Aibo. `http://www.sonyaibo.net/aboutaibo.html`, 2004. Last accessed on 2004.

[12] R.C. Arkin, M. Fujita, T. Takagi, and R. Hasegawa. An ethological and emotional basis for human–robot interaction. *Robotics and Autonomous Systems*, 42(3):191–201, 2003.

[13] Probo. `http://probo.vub.ac.be/`, 2009. Last accessed on 2009.

[14] J. Saldien, K. Goris, B. Vanderborght, and D. Lefeber. On the design of an emotional interface for the huggable robot probo. In *AISB Symposium*, pages 1–6, 2008.

[15] Papero. `http://www.nec.co.jp/products/robot/en/index.html`.

[16] J. Osada, S. Ohnaka, and M. Sato. The scenario and design process of childcare robot, papero. In *International Conference on Advances in Computer Entertainment Technology*, 2006.

[17] C. Jones and A. Deeming. Affective human-robotic interaction. *Affect and Emotion in Human-Computer Interaction*, pages 175–185, 2008.

[18] M. Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[19] J.D. Gould and C. Lewis. Designing for usability: key principles and what designers think. *Communications of the ACM*, 28(3):300–311, 1985.

[20] K. Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds. Socially intelligent agents. *Socially Intelligent Agents*, pages 1–20, 2002.

[21] P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, 1987.

[22] RD Walk and KL Walters. Perception of the smile and other emotions of the body and face at different distances. *Bulletin of the Psychonomic Society*, 26(6):510–510, 1988.

[23] R.A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[24] P.N. Juslin and K.R. Scherer. *Vocal expression of affect*. Oxford University Press, Oxford, UK, 2005.

[25] W. Stiehl and C. Breazeal. Affective touch for robotic companions. *Affective Computing and Intelligent Interaction*, pages 747–754, 2005.

[26] A. Mehrabian. Communication without words. *Psychological Today*, 2:53–55, 1968.

[27] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187, 2007.

[28] Y. Gao, M.K.H. Leung, S.C. Hui, and M.W. Tananda. Facial expression recognition from line-based caricatures. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 33(3):407–412, 2003.

[29] C. Shan, S. Gong, and P.W. McOwan. A comprehensive empirical study on linear subspace methods for facial expression analysis. In *International Conference on Computer Vision and Pattern Recognition Workshop*, pages 153–153, 2006.

[30] S. Zafeiriou and I. Pitas. Discriminant graph structures for facial expression recognition. *IEEE Transactions on Multimedia*, 10(8):1528–1540, 2008.

[31] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[32] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[33] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[34] E. Kokiopoulou and Y. Saad. Orthogonal neighborhood preserving projections. In *International Conference on Data Mining*, pages 1–8, 2005.

[35] K. Dautenhahn, B. Ogden, and T. Quick. From embodied to socially embedded agents–implications for interaction-aware robots. *Cognitive Systems Research*, 3(3):397–428, 2002.

[36] B.M. Scassellati. *Foundations for a theory of mind for a humanoid robot*. PhD thesis, Massachusetts Institute of Technology, 2001.

[37] B. Adams, C. Breazeal, R.A. Brooks, and B. Scassellati. Humanoid robots: a new kind of tool. *Intelligent Systems and Their Applications*, 15(4):25–31, 2000.

[38] H.J. Ryu, SS Kwak, and M.S. Kim. A study on external form design factors for robots as elementary school teaching assistants. In *International Symposium on Robot and Human Interactive Communication*, pages 1046–1051, 2007.

[39] T. Ishida. Development of a small biped entertainment robot qrio. In *International Symposium on Micro-Nanomechatronics and Human Science*, pages 23–28, 2004.

[40] F. Tanaka, A. Cicourel, and J.R. Movellan. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, 104(46):17954–17958, 2007.

[41] J.R. Movellan, F. Tanaka, I.R. Fasel, C. Taylor, P. Ruvolo, and M. Eckhardt. The rubi project: a progress report. In *International Conference on Human-Robot Interaction*, pages 333–339, 2007.

[42] P. Ruvolo, I. Fasel, and J. Movellan. Auditory mood detection for social and educational robots. In *International Conference on Robotics and Automation*, pages 3551–3556, 2008.

[43] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *International Conference on Automatic Face and Gesture Recognition*, pages 223–230, 2006.

[44] E. Hyun and H. Yoon. Characteristics of young children's utilization of a robot during play time: A case study. In *International Symposium on Robot and Human Interactive Communication*, pages 675–680, 2009.

[45] K. Wada and T. Shibata. Living with seal robotsits sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, 23(5):972–980, 2007.

[46] T. Shibata, T. Mitsui, K. Wada, A. Touda, T. Kumasaka, K. Tagami, and K. Tanie. Mental commit robot and its application to therapy of children. In *International Conference on Advanced Intelligent Mechatronics*, volume 2, pages 1053–1058, 2001.

[47] H. Kozima, M.P. Michalowski, and C. Nakagawa. Keepon a playful robot for research, therapy, and entertainment. *International Journal of Social Robotics*, 1(1):3–18, 2009.

[48] M. Poel, D. Heylen, A. Nijholt, M. Meulemans, and A. Van Breemen. Gaze behaviour, believability, likability and the icat. *AI & Society*, 24(1):61–73, 2009.

[49] S. Yun, J. Shin, D. Kim, C.G. Kim, M. Kim, and M.T. Choi. Engkey: Tele-education robot. In *International Conference on Social Robotics*, volume 7072, pages 142–152, 2011.

[50] P. Marti and L. Giusti. A robot companion for inclusive games: A user-centred design perspective. In *International Conference on Robotics and Automation*, pages 4348–4353, 2010.

[51] C. Darwin. The expression of the emotions in man and animals, 1872.

[52] P. Ekman and W.V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[53] P. Ekman and W.V. Friesen. Facial action coding system: A technique for the measurement of facial movement, 1978.

[54] Emfacs – scoring for emotion with facs. `http://face-and-emotion.com/dataface/facs/emfacs.jsp`, 2003.

[55] Facial action coding system affect interpretation dictionary (facsaid). `http://face-and-emotion.com/dataface/facsaid/description.jsp`, 2003.

[56] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[57] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.

[58] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[59] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[60] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[61] C. Shan, S. Gong, and P.W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[62] G. Zhao and M. Pietikäinen. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern Recognition Letters*, 30(12):1117–1127, 2009.

[63] P. Yang, Q. Liu, and D.N. Metaxas. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30(2):132–139, 2009.

[64] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[65] Y.L. Tian, T. Kanade, and J.F. Cohn. Facial expression analysis. *Handbook of Face Recognition*, pages 247–275, 2005.

[66] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2010.

[67] Y. Su, Y. Fu, Q. Tian, and X. Gao. Cross-database age estimation based on transfer learning. In *International Conference on Acoustics Speech and Signal Processing*, 2010.

[68] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning. Local features based facial expression recognition with face registration errors. In *International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2008.

[69] A. Lanitis, C.J. Taylor, and T.F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[70] M.J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.

[71] D. DeCarlo and D. Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In

International Conference on Computer Vision and Pattern Recognition, pages 231–238, 1996.

[72] Y.I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[73] T. Moriyama, T. Kanade, J.F. Cohn, J. Xiao, Z. Ambadar, J. Gao, and H. Imamura. Automatic recognition of eye blinking in spontaneously occurring behavior. In *International Conference on Pattern Recognition*, volume 4, pages 78–81, 2002.

[74] M.S. Bartlett, B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T.J. Sejnowski, and J.R. Movellan. Automatic analysis of spontaneous facial behavior: A final project report. *University of California at San Diego*, 2001.

[75] M.W. Sullivan and M Lewis. Emotional expressions of young infants and children. *Infants and Young Children*, 16(2):120–142.

[76] J. Cassell. Towards a model of technology and literacy development: Story listening systems. *Journal of Applied Developmental Psychology*, 25(1):75–105, 2004.

[77] I. Verenikina, P. Harris, and P. Lysaght. Child's play: computer games, theories of play and children's development. In *ACM International Conference Proceeding Series*, volume 98, pages 99–106, 2003.

[78] Limit switch. `http://en.wikipedia.org/wiki/Limit_switch`. Last accessed on 2012.

[79] Program - aimlbot.dll. `http://aimlbot.sourceforge.net/`, 2006.

[80] Socketcoder.com. http://www.socketcoder.com/ArticleFile.aspx?index=2&ArticleID=72, 2011.

[81] Sending and playing microphone audio over network. http://www.codeproject.com/Articles/19854/Sending-and-playing-microphone-audio-over-network, 2006.

[82] L. Wang, Y. Zhang, and J. Feng. On the euclidean distance of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1334–1339, 2005.

[83] M.J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[84] J. Lu and Y.P. Tan. A doubly weighted approach for appearance-based subspace learning methods. *IEEE Transactions on Information Forensics and Security*, 5(1):71–81, 2010.

[85] Y. Fu, S. Yan, and T.S. Huang. Correlation metric for generalized feature extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2229–2235, 2008.

[86] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[87] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[88] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998.

[89] F. Wallhoff. Facial expressions and emotion database. *Technische Universität München*, 2006.

[90] M.A. Goodrich and A.C. Schultz. Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2007.

[91] A. Niculescu, B. van Dijk, A. Nijholt, S.L. Swee, and H. Li. How humans behave and evaluate a social robot in real-environment settings. In *Proceedings of the Annual European Conference on Cognitive Ergonomics*, pages 351–352, 2010.

Appendix: Publications arising from this PhD work

1. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Adaptive discriminative metric learning for facial expression recognition", IET Biometrics, vol. 1, no. 3, pp. 160-167, 2012.

2. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "A survey on perception methods for human-robot interaction in social robots", International Journal of Social Robotics, under minor revision.

3. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Misalignment-robust subspace analysis for facial expression recognition," submitted to International Journal of Pattern Recognition and Artificial Intelligence, under review.

4. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Dorothy Robotubby: A Robotic Nanny", International Conference on Social Robotics, pp. 118-127, 2012.

5. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Cross-dataset facial expression recognition", IEEE International Conference on Robotics and Automation, pp. 5985-5990, 2011.

6. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Weighted biased linear

discriminant analysis for misalignment-robust facial expression recognition", IEEE International Conference on Robotics and Automation, pp. 3881-3886, 2011.

7. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Enhanced projection functions for eye localization under controlled conditions", International Universal Communication Symposium, 2011.

8. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Exploring feature descriptors for misalignment-robust facial expression recognition", International Conference on Humanoid, Nanotechnology, Information Technology Communication and Control, Environment and Management, 2011.

9. **Haibin Yan**, Marcelo H. Ang Jr, Aun Neow Poo, "Misalignment robust facial expression recognition", 4th Asia International Symposium on Mechatronics, pp. 105-110, 2010.