Chapter 2
Body Movement Analysis and Recognition

Yang Xiao, Hui Liang, Junsong Yuan and Daniel Thalmann

Abstract In this chapter, a nonverbal way of communication for human–robot interaction by understanding human upper body gestures will be addressed. The human–robot interaction system based on a novel combination of sensors is proposed. It allows one person to interact with a humanoid social robot with natural body language. The robot can understand the meaning of human upper body gestures and express itself by using a combination of body movements, facial expressions, and verbal language. A set of 12 upper body gestures is involved for communication. Human–object interactions are also included in these gestures. The gestures can be characterized by the head, arm, and hand posture information. CyberGlove II is employed to capture the hand posture. This feature is combined with the head and arm posture information captured from Microsoft Kinect. This is a new sensor solution for human-gesture capture. Based on the body posture data, an effective and real-time human gesture recognition method is proposed. For experiments, a human body gesture dataset was built. The experimental results demonstrate the effectiveness and efficiency of the proposed approach.

2.1 Introduction

Recently, human–robot interaction (HRI) has drawn the attention of the academic and industrial communities. Regarded as the sister community of human–computer interaction (HCI), HRI is still a relatively young field that began to emerge in the

Y. Xiao (✉)
School of Automation, Huazhong University of Science and Technology, Wuhan, China
e-mail: Yang_Xiao@hust.edu.cn

H. Liang · J. Yuan · D. Thalmann
BeingThere Centre, Nanyang Technological University, Singapore, Singapore

J. Yuan
e-mail: JSYUAN@ntu.edu.sg

D. Thalmann
e-mail: danielthalmann@ntu.edu.sg

N. Magnenat-Thalmann et al. (eds.), *Context Aware Human-Robot and Human-Agent Interaction*, Human–Computer Interaction Series,
DOI 10.1007/978-3-319-19947-4_2

Fig. 2.1 Human–robot social interaction, with human on the *right* and robot on the *left*

1990s [10, 14]. It is an interdisciplinary research field that requires contributions from mathematics, psychology, mechanical engineering, biology, computer science, etc. [14].

HRI aims to understand and shape the interactions between humans and robots. Unlike early interactions, more social dimensions must be considered in HRI, especially when interactive social robots are involved [10, 13]. In this case, robots should be believable. Moreover, humans prefer to interact with robots as they do with other people [10, 13]. Therefore, one way to increase believability would be to make the robot interact with humans using the same modalities as human–human interaction. This includes verbal and body language as well as facial expressions; i.e., the robots should be able to use these modalities for both perception and expression. Some social robots have already been proposed toward this goal. For instance, the Leonardo robot expresses itself using a combination of voice, facial, and body expressions [25]. Another example is the Nao humanoid robot[1] that can use vision along with gestures and body expression of emotions [2]. Different from these two robots, the Nadine robot is a highly realistic humanoid robot (Fig. 2.1). This robot presents some different social challenges. In this chapter, a human–robot interaction system that addresses some of these challenges is proposed. As shown in Fig. 2.1, it supports a person to communicate and interact with a humanoid robot. In the proposed system, the human can naturally communicate with the Nadine robot using body language. The Nadine robot is able to express herself by using a combination of speech, body language, and facial expressions. In this chapter, the main research concern addressed is how to establish communication between human and robot using body language.

Verbal and nonverbal language are two means of communication for human–human interaction. Verbal language has been used in many HRI systems

[1] http://www.aldebaran-robotics.com/

Fig. 2.2 Human–human interaction accompanied with nonverbal language

[11, 21, 23, 26–28]; however, it still has some constraints. That is, speech recognition accuracy is likely to be affected by the background noise, human accents, and device performance. Moreover, learning and interpreting the subtle rules of syntax and grammar in speech is a difficult task. These factors limit the practical use of verbal language to a certain degree. On the other hand, nonverbal clues can convey rich communication messages [7, 19]. Evidently, they play an important role in human–human interaction to reinforce the communication performance as shown in Fig. 2.2. Thus, one of our research motivations is *to apply nonverbal language to human–robot social interaction.* More specifically, upper body gesture language is employed. Currently, 12 human upper body gestures are involved in the proposed system, which are all natural gestures with intuitive semantics. They are characterized by head, arm, and hand posture information simultaneously. It is worth noting that *human–object* interactions are involved in these gestures. *Human–object* interaction events manifest frequently during human–human interaction in daily life. However, to our knowledge, they were largely ignored by previous HRI systems.

The main challenge to apply upper body gesture language to human–robot interaction is how to enable the Nadine robot to understand and react to human gestures accurately and in real-time. To achieve this goal, two crucial issues need to be solved:

- First, an appropriate human gesture-capture sensor solution is required. To recognize the 12 upper body gestures, head, arm, and hand posture information is needed simultaneously. As robustly obtaining hand posture based on vision-based sensors (such as the RGB camera) is still a difficult task [18, 29], the wearable CyberGlove II [15] (shown in Fig. 2.3) is used. Using this device, high-accuracy hand

Fig. 2.3 CyberGlove II

Fig. 2.4 The microsoft kinect RGB and depth sensor

posture data can be acquired stably. Meanwhile, Microsoft Kinect [24] (shown in Fig. 2.4) is an effective and efficient low-cost depth sensor applied successfully to human body tracking. The skeletal joints can be extracted from the Kinect depth images [24] in real-time (30 fps). In our work, Kinect is applied to capture the upper body (head and arm) posture information. Recently, Kinect 2 that supports tracking multiple people with better depth imaging quality was released. Since our work investigates the HRI scenario involving only one person, Kinect is sufficient to handle the human body tracking task;

- Second, an effective and real-time gesture recognition method should be developed. Based on the CyberGlove II and Kinect posture data, descriptive upper body gesture feature is proposed. To leverage the gesture understanding performance, LMNN distance metric learning method [33] is applied. Then, the energy-based LMNN classifier is used to recognize the gestures.

To evaluate the proposed gesture recognition method, a human upper body gesture dataset is constructed. This dataset contains gesture samples from 25 people of different genders, body sizes, and culture backgrounds. The experimental results demonstrate the effectiveness and efficiency of our method.

Overall, the main contributions of this chapter include:

- A novel human gesture-capture sensor solution is proposed. That is, the CyberGlove II and Kinect are integrated to capture head, arm, and hand posture information simultaneously;
- An effective and real-time upper body gesture recognition approach is proposed;
- To support humans to communicate and interact with robots using body language, a gesture understanding and human–robot interaction (GUHRI) system is built.

The remainder of this chapter is organized as follows. Problematic issues are discussed in Sect. 2.2. Section 2.3 gives an overview of the related state-of-the-art works. The recent approaches are described in Sect. 2.4. Section 2.5 introduces the future avenues. The conclusions are drawn in Sect. 2.6.

2.2 Problematic

To successfully apply nonverbal language to HRI by understanding human upper body gestures, some critical problematic issues and challenges need to be addressed. A brief discussion of this point has been made in Sect. 2.1. In this section, we extend the discussion from the perspectives of the HRI system in detail.

- As our research aims to promote social interaction between humans and robot, the semantics of natural human body language need to be analyzed. To fully understand the human upper body gestures, head, arm, and hand posture information should be captured simultaneously in an effective way. This proposition is effectively demonstrated by Figs. 2.1 and 2.2. However, to our knowledge, very few previous body gesture recognition works take hand and rough body posture information into consideration together. Thus, appropriate human gesture-capture devices are the essential components to construct a successful HRI system. For human gesture capture, vision-based sensors are the trend as they exert no burden on the users and lead to better user experience. Some successful examples have already emerged recently, such as Microsoft Kinect, which is applied to human body parsing and tracking. However, under unconstrained conditions, it is still difficult to capture the hand posture robustly because of drastic hand rotation and serious occlusion as shown in Fig. 2.5. Meanwhile, it is not feasible to restrict the user's hand position and orientation during the phase of natural HRI. As a consequence, according to the current capacity of vision-based sensors, they are not the optimal choice to capture the hand posture for the HRI system. Thus, a more applicable human gesture-capture sensor solution should be proposed. In addition to effectiveness,

Fig. 2.5 The different hand gestures with drastic rotation and serious occlusion

another important factor for human gesture capture is efficiency. As HRI is in high real-time demand, the data acquisition stage must be finished as soon as possible.

- Based on the reliable hand and body posture feature, how to recognize the upper body gestures effectively in real-time is the latest concern that needs to be addressed. Since the hand and rough body posture information is captured from different sensors, i.e., they are multimodular data, how to fuse them to form a unified body gesture description is the first point we focus on. Second, for HRI application an adequate classification scheme should be proposed to leverage the performance, including the distance metric learning method and the choice of classifier. Last but not least, enough training samples are required to drive the supervised body gesture recognition approach. However, there is no existing body gesture dataset that can be applied to our work directly. Thus, building a novel upper body gesture dataset with sufficient available samples is another crucial task.
- Whether robots can naturally react to human body language will largely affect the user experience. Since the Nadine robot has the capacity to express herself using a combination of speech, body language, and facial expressions, a suitable interaction scenario is required for robot control to make her more humanlike and vivid. Indeed, the interaction scenario should be designed according to the human habits during human–human interaction.

2.3 State of the Art

HRI systems are constructed mainly based on verbal, nonverbal, or multimodal communication modalities. As mentioned in Sect. 2.1, verbal language still faces constraints in practical applications. Our work focuses on studying how to apply nonverbal language to human–robot social interaction, especially using upper body gesture language. Some HRI systems have already employed body gesture language for human–robot communication. In [30], an arm gesture-based interface for HRI was proposed. The user could control a mobile robot using static or dynamic arm gestures. Hand gesture was used as the communication modality for HRI in [5]. The HRI systems addressed in [27, 28] could recognize the human pointing gesture using the 3D head and hand position information, and head orientation was further appended to leverage the performance. In [11], the social robot could understand the human body language characterized by arm and head posture. Our proposition on nonverbal human–robot communication is different from previous works mainly in two aspects. First, head, arm, and hand posture are jointly captured to describe the 12 upper body gestures involved in the GUHRI system. Second, the gestures accompanied with human–object interaction can be understood by the robot. These gestures were always ignored by previous HRI systems, although they manifest frequently in daily life as shown in Fig. 2.6.

Fig. 2.6 Human–object interactions in daily life

Body gesture recognition plays an important role in the GUHRI system. According to the gesture-capture sensor type, gesture recognition systems can be categorized as encumbered and unencumbered [4]. Encumbered systems require the user to wear physical assistive devices such as infrared responders, hand markers, or data gloves. These systems have high precision and fast response, and are robust to environmental changes. Many encumbered systems have been proposed. For instance, two education systems [1] were built for the deaf using data gloves and optical motion capture devices; Lu et al. [18] proposed an immersive virtual object manipulation system based on two data gloves and a hybrid ultrasonic tracking system. Although most commercialized gesture-capture devices are currently encumbered, unencumbered systems are expected to be the future choice, especially vision-based systems. With the emergence of low-cost 3D vision sensors, the application of such devices becomes a hot topic in both the research and commercial fields. One of the most famous examples is Microsoft Kinect, which has been successfully employed in human body tracking [24], activity analysis [31], and gesture understanding [37, 38]. Even for other vision applications (such as scene categorization [36] and image segmentation [34, 35]), Kinect holds the potential to boost the performance. However, accurate and robust hand posture capture is still a difficult task for vision-based sensors.

As discussed above, both encumbered and unencumbered sensors possess intrinsic advantages and drawbacks. For specific applications, they can be complementary. In the GUHRI system a tradeoff between the two kinds of sensors is made, that is, the fine hand posture is captured by the encumbered device (CyberGlove II), while the rough upper body posture is handled by the unencumbered sensor (Kinect).

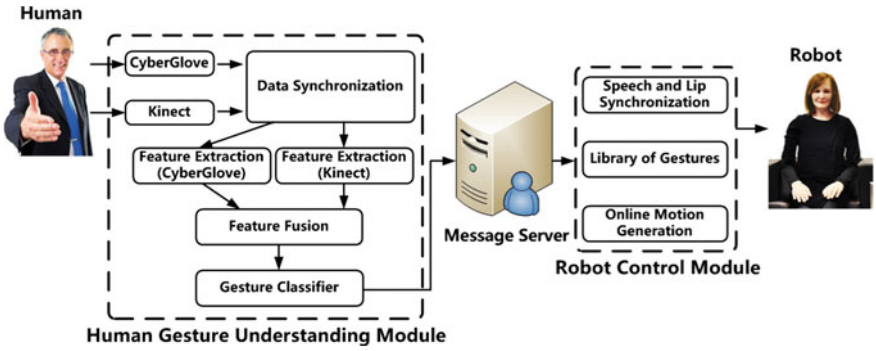

Fig. 2.7 The GUHRI system architecture

2.4 Recent Approaches

In this section, we give an overview of the GURHI system. The human upper body gesture understanding method is then illustrated. Next, the scenario for HRI is introduced. Experiment and discussion are finally given.

2.4.1 System Overview

The proposed GUHRI system is able to capture and understand human upper body gestures and trigger the robot's reaction in real-time accordingly. The GUHRI system is implemented using a framework called Integrated Integration Platform (I2P) specifically developed for integration. I2P was developed by the Institute for Media Innovation.[2] This framework allows for the link and integration of perception, decision, and action modules within a unified and modular framework. The platform uses client–server communications between the different components. Each component has an I2P interface and communication between the client and servers is implemented using thrift.[3] It should be noted that the framework is highly modular and components can be added to make the GUHRI system extendable. As shown in Fig. 2.7, the current GUHRI system is mainly composed of two modules. One, the *human gesture understanding module* that serves as the communication interface between human and robot, and the other is the *robot control module* proposed to control the robot's behaviors for interaction. At this stage, our system supports the interaction between one person and one robot.

One right-handed CyberGlove II and one Microsoft Kinect are employed to capture the human hand and body posture information simultaneously for gesture

[2] http://imi.ntu.edu.sg/Pages/Home.aspx

[3] http://thrift.apache.org/

understanding. This is a new gesture capturing sensor solution, different from all the approaches introduced in Sect. 2.3. Specifically, CyberGlove is used to capture the hand posture, and Kinect is applied to acquire the 3D position information of the human skeletal joints (including head, shoulder, limb, and hand). At this stage, the GUHRI system relies on the upper body gestures triggered by the human right hand and right arm.

Apart from the CyberGlove, the user does not need to wear any other device. Thus, the proposed sensor solution does not exert a heavy burden to make the user uncomfortable. Meanwhile, as the CyberGlove II is involved in the system using Bluetooth, the user can move freely. In addition, the GUHRI system is able to recognize gestures with human–object interaction, such as "*call*", "*drink*", "*read*" and "*write*" by fusing the hand and body posture information. These gestures were often ignored by previous systems. However, they manifest frequently during the daily interaction between humans. These affect the interaction state abruptly and should be considered as essential elements of the natural HRI. In our system, the robot is able to recognize and give meaningful responses to these gestures (Fig. 2.8).

The first step of the gesture understanding phase is to synchronize the original data from CyberGlove and Kinect. The descriptive features are then extracted from them respectively. The multimodal features are then fused to generate the unified input for the gesture classifier. Lastly, the gesture recognition and understanding results are sent to the robot control module via message server to trigger the robot's reaction.

The robot control module enables the robot to respond to the human's body gesture language. In our system, the robot's behavior is composed of three parts: body movement, facial expression, and verbal language. Combining these modalities makes the robot more lifelike, and should enhance the users' interest during interaction.

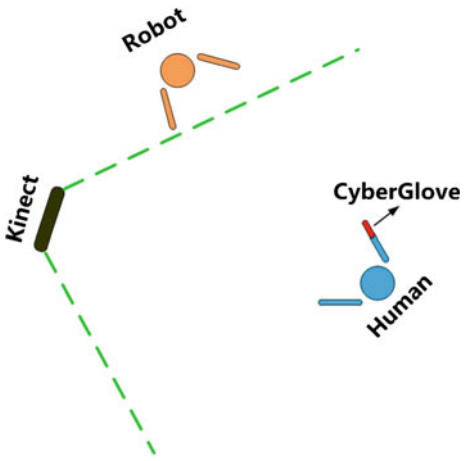

Fig. 2.8 The GUHRI system deployment

2.4.2 Human Upper Body Gesture Understanding

As an essential part of the GUHRI system, the human upper body gesture understanding module plays an important role during interaction. Its performance will highly affect the interaction experience. In this section, our upper body gesture understanding method by fusing the gesture information from CyberGlove and Kinect is illustrated in detail. First, the body gestures included in the GUHRI system are introduced. The feature extraction pipelines for both CyberGlove and Kinect are then presented. To generate an integral gesture description, the multimodal features from different sensors are fused as the input for classifier. Aiming to enhance the gesture recognition accuracy, LMNN distance metric learning approach [33] is applied for mining the optimal distance measures, and the energy-based classifier [33] is applied for decision making.

2.4.2.1 Gestures in the GUHRI System

At the current stage, 12 static upper body gestures are included in the GUHRI system. As we only have one right-hand CyberGlove, to obtain accurate hand posture information all the gestures are mainly triggered by the human right hand and right arm. The involved gestures can be partitioned into two categories, according to whether human–object interaction happens:

- *Category 1*: body gestures *without* human–object interaction;
- *Category 2*: body gestures *with* human–object interaction.

Category 1 contains 8 upper body gestures: "*be confident*", "*have question*", "*object*", "*praise*", "*stop*", "*succeed*", "*shake hand*" and "*weakly agree*." Some gesture samples are shown in Fig. 2.9. These gestures are natural and have intuitive meaning. They are related to the human's emotional state and behavior intention and are not ad hoc for specific applications. Therefore, gesture-to-meaning mapping is not needed in our system. As human behavior habits are not all the same, recognizing natural gestures is more challenging than ad hoc ones. However, natural gestures are more meaningful for HRI. As shown in Fig. 2.9, both hand and body posture information are required for recognizing these gestures. For instance, the upper body postures corresponding to "*have question*" and "*object*" are very similar. Without the hand posture, they are difficult to distinguish. The same happens to "*have question*," "*weakly agree*," and "*stop*." That is, they correspond to similar hand gestures but very different upper body postures.

Category 2 is composed of four other upper body gestures: "*call*," "*drink*," "*read*," and "*write*" (Fig. 2.10). Being different from *Category 1* gestures, these four gestures happen with human–object interactions. Existing systems do not consider such gestures (see Sect. 2.3). One main reason is that objects often cause body occlusion, especially to the hand. In this case, vision-based hand gesture recognition methods are impaired. Hence, the CyberGlove is employed to capture the hand posture. In the

Fig. 2.9 The *Category 1 upper body* gestures. These gestures can be characterized by the body and hand posture information simultaneously

Fig. 2.10 The *Category 2 upper body* gestures. These gestures can be characterized by the body and hand posture information simultaneously

GUHRI system, *Category 2* gestures are recognized and affect the interaction in a realistic way. These gestures are also recognized based on the hand and upper body posture information.

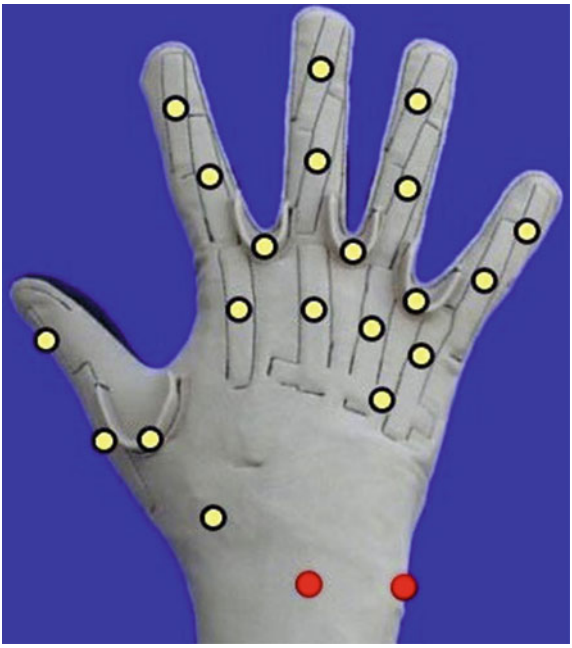

Fig. 2.11 CyberGlove II data joints [15]

2.4.2.2 Feature Extraction and Fusion

In this section, we introduce the feature extraction methods for both human hand and upper body posture description. The multimodal feature fusion approach is also illustrated.

(a) Hand Posture Feature

As discussed above, the description of the human hand and upper body posture is key to recognize and understand the 12 upper body gestures.

The immersion wireless CyberGlove II is used as the hand posture capture device in the GUHRI system. As one of the most sophisticated and accurate data gloves, CyberGlove II provides 22 high-accuracy joint-angle measurements in real-time. These measurements reflect the bending degree of the fingers and wrist. The 22 data joints (marked as big white or black dots) are located on the CyberGlove as shown in Fig. 2.11. However, not all the joints are used. For the hand gestures in our application, we found that the wrist posture does not provide stable descriptive information. The wrist bending degrees of different people vary to a large extent even for the same gesture. This phenomenon is related to different behavior habits. This is the reason that the two wrist data joints (marked as black) are discarded. A 20-dimensional feature vector F_{hand} is extracted from the 20 white data joints to describe the human hand posture as

$$F_{\text{hand}} = (h_1, h_2, h_3 \ldots h_{19}, h_{20}), \tag{2.1}$$

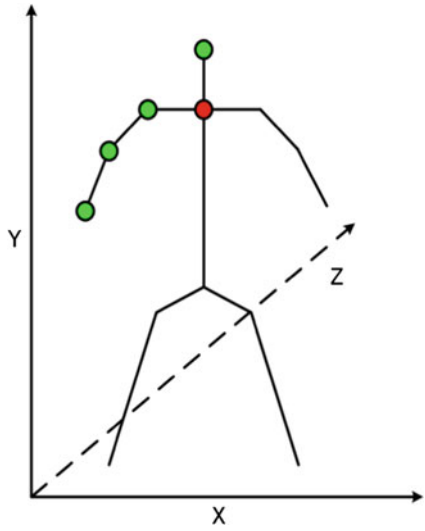

Fig. 2.12 The selected body skeletal joints

where h_i is the bending degree corresponding to the white data joint i.

(b) Upper Body Posture Feature

Using the Kinect sensor, we shape the human upper body posture intermediately using the 3D skeletal joint positions. For a full human subject, 20 body joint positions can be detected and tracked by the real-time skeleton tracker [24] based on the Kinect depth frame. This is invariant to posture, body shape, clothing, etc. Each joint J_i is represented by three coordinates at the frame t as

$$J_i = (x_i(t), y_i(t), z_i(t)). \tag{2.2}$$

However, not all the 20 joints are necessary for upper body gesture recognition. As aforementioned, head and right arm are highly correlated with the 12 upper body gestures (Figs. 2.9 and 2.10). For efficiency, only four upper body joints are chosen as the descriptive joints for gesture understanding. These are "*head*," "*right shoulder*," "*right elbow*," and "*right hand*" that are shown as gray dots in Fig. 2.12.

Directly using the original 3D joint information for body posture description is not stable, because it is sensitive to the relative position between human and Kinect. Solving this problem by restricting the human's position is not appropriate for interaction. In [31], human action is recognized by using the pairwise relative positions between all joints, which is robust to the human–Kinect relative position. Inspired by this work, a simplified solution is proposed. First, the "*middle of the two shoulders*" joint (black dot in Fig. 2.12) is selected as the reference joint. The pairwise relative positions between the four descriptive joints and the reference joint are then computed for body posture description as

$$J_{sr} = J_s - J_r, \tag{2.3}$$

where J_s is the descriptive joint and J_r is the reference joint. With this processing, J_{sr} is less sensitive to the human–Kinect relative position. It is mainly determined by the body posture. The "*middle of the two shoulders*" was chosen as the reference joint because it can be robustly detected and tracked in most cases. Moreover, it is rarely occluded by the limbs or the objects when the gestures in GUHRI system happen. Finally, an upper body posture feature vector F_{body} of 12 dimensions is constructed by combining the four pairwise relative positions as

$$F_{\text{body}} = (J_{1r}, J_{2r}, J_{3r}, J_{4r}), \tag{2.4}$$

where J_{1r}, J_{2r}, J_{3r} and J_{4r} are the pairwise relative positions.

(c) Feature Fusion

From CyberGlove II and Kinect, two multimodal feature vectors: F_{hand} and F_{body} are extracted to describe the hand posture and upper body posture respectively. To fully understand the upper body gestures, the joint information about the two feature vectors is required. Both are essential for the recognition task. However, the two feature vectors locate in different value ranges. Simply combining them as the input for classifier will yield performance bias on the feature vector of low values. To overcome this difficulty, we scale them into similar ranges before feature fusion. Suppose F_i is one dimension of F_{hand} or F_{body}, $F_i^{\max}$ and $F_i^{\min}$ are the corresponding maximum and minimum value in the training set. Then F_i can be normalized as

$$\hat{F}_i = \frac{F_i - F_i^{\min}}{F_i^{\max} - F_i^{\min}}, \tag{2.5}$$

for both training and test.

After normalization, the effectiveness of the two feature vectors for gesture recognition will be balanced. Finally, they are fused to generate an integral feature vector by concatenation as

$$\mathbf{F} = (\hat{F}_{\text{hand}}, \hat{F}_{\text{body}}). \tag{2.6}$$

This process results in a 32-dimensional feature vector $\mathbf{F}$ used for upper body gesture recognition.

2.4.2.3 Classification Method

Using $\mathbf{F}$ as the input feature, the upper body gestures will be recognized by template matching based on the *energy-based LMNN classifier* proposed in [33].[4] It is derived from the energy-based model [8] and the LMNN distance metric learning method [33]. The latter part is the key to constructing this classifier. LMNN distance metric learning approach is proposed to seek the best distance measure for the

[4]The source code is available at http://www.cse.wustl.edu/~kilian/code/lmnn/lmnn.html

k-nearest neighbor (KNN) classification rule [9]. As one of the oldest methods for pattern recognition, KNN classifier is simple to implement and use. Nevertheless, it can still yield comparative results in certain domains such as object recognition and shape matching [3], it has also been applied to action recognition [20].

The KNN rule classifies each testing sample by the majority label voting among its k-nearest training samples. Its performance crucially depends on how to compute the distances between different samples for the k nearest neighbors search. Euclidean distance is the most widely used distance measure, although it ignores any statistical regularities that may be estimated from the training set. Ideally, the distance measure should be adjusted according to the specific task being solved. To achieve better classification performance, LMNN distance metric learning method is proposed to mine the best distance measure for the KNN classification.

Let $\{(\mathbf{x}_i, y_i)\}_{i=n}^{n}$ be a training set of n labeled samples with inputs $\mathbf{x}_i \in \mathbb{R}^d$ and class labels y_i. The main goal of LMNN distance metric learning is to learn a linear transformation $\mathbf{L} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that is used to compute the square sample distances as

$$\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2. \tag{2.7}$$

Using $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j)$ as the distance measure tends to optimize the KNN classification by making each input $\mathbf{x}_i$ have k nearest neighbors that share the same class label y_i to the greatest possibility. Figure 2.13 gives an intuitive illustration on LMNN distance metric learning. Compared with Euclidean distance, LMNN distance tries to pull the nearest neighbors of class y_i closer to $\mathbf{x}_i$, while pushing the neighbors from different classes away. On the assumption that the training set and the test set keep the similar feature distribution, LMNN distance metric learning can help to improve the KNN classification result.

The energy-based LMNN classifier makes use of both the $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j)$ distance measure and the loss function defined for LMNN distance metric learning. It constructs an energy-based criterion function, and the testing sample is assigned to the class that yields the minimum loss value. As the related theory is sophisticated, we do not give a detailed definition of the energy-based LMNN classifier here; readers can turn to [33] for reference.

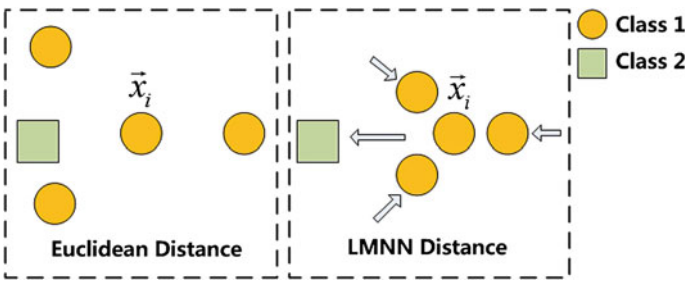

Fig. 2.13 Illustration of the LMNN distance metric learning

Fig. 2.14 Examples of body movements and facial expressions from the library of gestures

2.4.2.4 Human–Robot Interaction

As a case study for the GUHRI system, a scenario was defined in which a user and the robot interact in a classroom. The robot is the lecturer, and the user is the student. The robot is a female named "Nadine." Nadine can understand the 12 upper body gestures described in Sect. 2.4.2.1 and react to the users' gestures accordingly. In our system, Nadine is humanlike and capable of reacting by combining body movement, facial expression, and verbal language. In this way, Nadine's reactions provide the user with vivid feedback. Figure 2.14 shows some examples of Nadine's body movements along with corresponding facial expressions. Nonverbal behaviors can help to structure the processing of verbal information as well as giving affective feedback during the interaction [6, 17]. Thus, body movements and facial expressions are expected to enhance the quality of the interaction with Nadine.

In this scenario, Nadine's behaviors are triggered by the users' body language. Her reactions are consistent with the defined scenario (see Table 2.1). Note that because it is difficult to fully describe the robot's body actions, the robot's movements and emotional display are described at a high level. All the 12 upper body gestures are involved. The GHURI system can also handle unexpected situations during the interaction. For example, Nadine can react appropriately even if the user suddenly answers a coming phone call.

2.4.2.5 Experiment and Discussion

A human upper body gesture dataset was built to test the proposed gesture recognition method. This dataset involves all the 12 upper body gestures mentioned in Sect. 2.4.2.1. The samples are captured from 25 volunteers of different genders,

Table 2.1 The scenario for human–robot interaction

Human gestures	Nadine's response	
	Nonverbal	Verbal
"*be confident*"	Happy	It is great to see you so confident
"*have question*"	Moderate	What is your question?
"*object*"	Sad	Why do you disagree?
"*praise*"	Happy	Thank you for your praise
"*stop*"	Moderate	Why do you stop me?
"*succeed*"	Happy	Well done. You are successful
"*shake hand*"	Happy	Nice to meet you
"*weakly agree*"	Head nod	OK, we finally reach an agreement
"*call*"	Head shake	Please turn off your phone
"*drink*"	Moderate	You can have a drink. No problem
"*read*"	Moderate	Please, take your time and read it carefully
"*write*"	Moderate	If you need time for taking notes, I can slow my presentation

body sizes, and races. During the sample collection, no strict constraint was imposed on the people. They carried out the gestures based on their own habits. The user–Kinect relative position was also not strictly limited. For convenience, CyberGlove II was precalibrated for all the people with a standard calibration. Due to the dataset collection setup, large diversities may exist among the gesture samples from different people. This will yield challenges on body gesture recognition. Figure 2.15 exhibits parts of the *Category 1* and *Category 2* gesture samples ("*have question*," "*succeed*," "*call*," and "*drink*") captured from five people for comparison. For the sake of brevity, not all the gestures are shown. The five descriptive and reference skeletal joints proposed in Sect. 2.4.2.2 are marked as big dots in Fig. 2.15, and they are connected by the straight segments to shape the upper body posture intuitively. From the exhibited samples, we can observe that:

- For the different people, the listed body gestures can indeed be differentiated from the hand and upper body posture information, and the people execute the gestures differently to a certain degree. As aforementioned, this phenomenon leads to challenges on upper body gesture recognition;
- For different people and gestures, the five skeletal joints employed for gesture recognition can be tracked robustly, even when human–object interaction occurs. Generally, their resulting positions are accurate for gesture recognition. Meanwhile, CyberGlove II is a human-touch device that can capture the hand posture

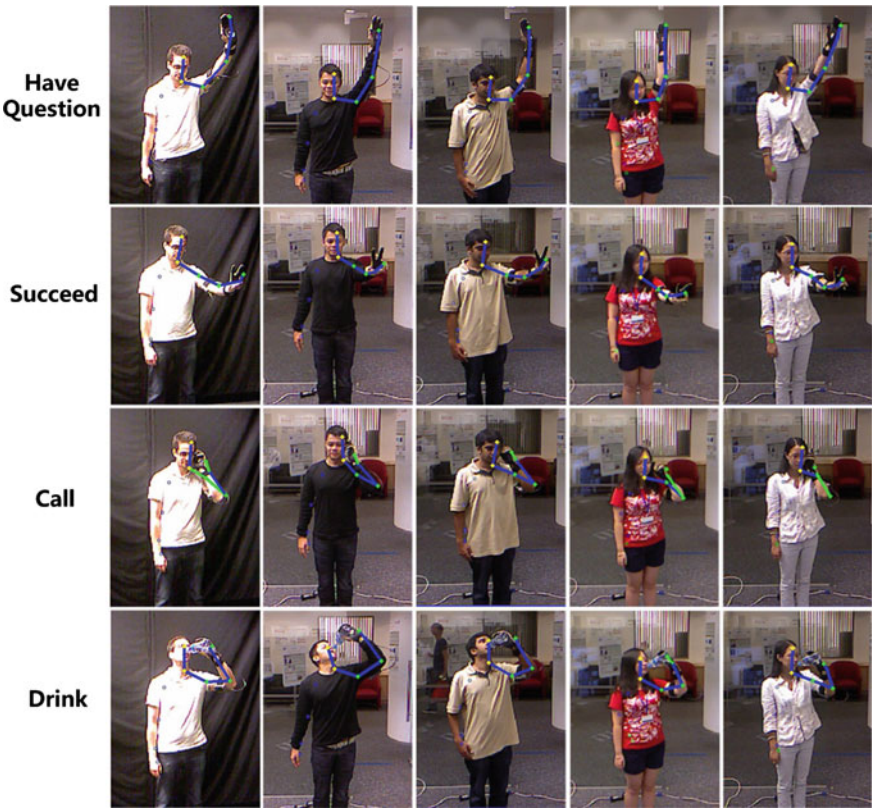

Fig. 2.15 Some gesture samples captured from different volunteers. These people are of different genders, body sizes, and races. They executed the gestures based on their own habits

robustly to yield high-accuracy data. Thus, the proposed human gesture-capture sensor solution can stably acquire available data for gesture recognition.

For each gesture, one key snapshot is picked up to build the dataset among all the 25 people. As a consequence, the resulting dataset contains $25 \times 12 = 300$ gesture samples in all. During experiment, the samples are randomly split into the training and testing set five times, and the average classification accuracy and standard deviation are reported.

The KNN classifier is used as the baseline to make comparison with the energy-based LMNN classifier. They are compared both on the items of classification accuracy and on time consumption. The KNN classifier runs with different kinds of distance measures. Following the experimental setup in [33],"k" is set as 3 in all cases. As the training sample number is a crucial factor that affects the classification accuracy, the results of two classifiers are compared corresponding to different amounts of training samples. For each class, the training sample number will increase from 4 to 14 with step size 2.

Table 2.2 Classification result (%) of the constructed *upper* body gesture dataset

Classifiers	Training sample number per class		
	4	6	8
KNN (Euclidean)	86.51(±2.89)	89.56(±2.43)	91.47(±2.21)
KNN (PCA)	73.81(±4.78)	84.04(±5.41)	79.31(±3.09)
KNN (LDA)	79.68(±5.33)	90.44(±2.56)	92.35(±2.44)
KNN (LMNN)	86.67(±2.18)	90.35(±2.56)	92.16(±1.80)
Energy (LMNN)	**90.00(±3.40)**	**92.28(±0.85)**	**94.31(±2.04)**
	10	12	14
KNN (Euclidean)	93.00(±1.15)	93.33(±0.73)	92.27(±2.30)
KNN (PCA)	86.11(±2.75)	88.21(±4.17)	86.67(±3.70)
KNN (LDA)	92.44(±1.34)	94.74(±0.84)	93.48(±1.38)
KNN (LMNN)	93.67(±1.34)	93.85(±1.48)	93.48(±1.15)
Energy (LMNN)	**95.22(±1.50)**	**95.64(±1.39)**	**96.52(±2.37)**

The best performance is shown in boldface. Standard deviations are in parentheses

Other two well-known distance metric learning methods, PCA [16] and LDA [12], are used for comparison with the LMNN distance metric learning approach. For PCA, the first 10 eigenvectors are used to capture roughly 90 % of the sum of eigenvalues, while the first 6 eigenvectors are used for LDA. The distance measures yielded by PCA and LDA are applied to the KNN classifier.

Table 2.2 lists the classification results yielded by the different classifier and distance measure combinations. It can be observed that:

- The 12 upper body gestures in the dataset can be well recognized by the proposed gesture recognition method. More than 95.00 % classification accuracy can be achieved if enough training samples are used. With the increase in training sample amount, the performance is generally enhanced consistently;
- Corresponding to all the training sample numbers, the energy-based LMNN classifier can yield the highest classification accuracy. Even with a small number (such as 4) of training samples it can still achieve relative good performance (90.00 %). When the training sample number reaches 14, the classification accuracy (96.52 %) is nearly satisfied for practical use, and its standard deviations are relatively low in most cases, which means that the energy-based LMNN classifier is robust to the gesture diversities among people;
- KNN classifier can also yield good results on this dataset. However, it is inferior to the energy-based LMNN classifier. Compared to Euclidean distance, LMNN distance metric learning method can improve the performance of KNN classifier consistently in most cases. However, it works much better on the energy-based model;
- PCA does not work well on this dataset. Its performance is worse than the basic Euclidean distance. The reason may be that PCA needs a large number of training

Table 2.3 Average testing time consumption (ms) per sample

Classifiers	Training sample number per class		
	4	6	8
KNN (LDA)	0.0317	0.0398	0.0414
KNN (LMNN)	0.0239	0.0282	0.0328
Energy (LMNN)	0.0959	0.1074	0.1273
	10	12	14
KNN (LDA)	0.0469	0.0498	0.0649
KNN (LMNN)	0.0342	0.0418	0.0525
Energy (LMNN)	0.1359	0.1610	0.1943

The program is run on the computer with Intel (R) Core (TM) i5-2430M @ 2.4 GHz (only using one core)

samples to obtain the satisfied distance measures [22]. This is the limitation for practical applications.

- LDA also achieves good performance for upper body gesture recognition. However, it is still consistently inferior to energy-based LMNN, especially when the training sample number is small. For example, when the training sample number is only 4, energy-based LMNN's accuracy (90.00 %) is significantly better than that of LDA (79.68 %) by a large margin (10.32 %).

Besides the classification accuracy, the testing time consumption is also what we are concerned about. The reason is that the GUHRI system should run in real-time for good HRI experience. According to the classification results in Table 2.2, the energy-based LMNN classifier, LMNN KNN classifier, and LDA KNN classifier are the three strongest ones for gesture recognition. Here, comparison on their testing time is also made. Table 2.3 lists the average running time per testing sample of the three classifiers, corresponding to different amounts of training samples. We can see that the three classifiers are extremely fast under our experimental conditions, and the time consumption mainly depends on the number of training samples. Frankly, the LDA KNN classifier and LMNN KNN classifier are much faster than the energy-based LMNN classifier. If a huge number of training samples were used (such as tens of thousands), the LDA KNN classifier and LMNN KNN classifier would be a better choice to achieve the balance between classification accuracy and computational efficiency.

2.5 Future Avenues

Our current research mainly pays attention to understanding the static upper body gestures. However, the dynamic ones are also the essential components of body language in daily life. For human–human interaction, they provide additional communication clues to boost the interaction performance. To make human–robot interaction

more natural and lifelike, the robot should be capable of understanding both static and dynamic gestures. Being different from static gestures, motion information is required to recognize the dynamic gestures. From Microsoft Kinect SDK, human body 3D joint information can be achieved in real-time (30 fps). Motion information can be intuitively extracted from the position change in body joints along the temporal axis. The body joint motion information has earlier been applied to activity recognition [31, 32]. Nevertheless, these works still ignore hand gestures, which may lead to ambiguity on gesture understanding. Thus, one of our future research avenues is to recognize the dynamic human upper body gestures (such as "*wave hand*," "*say no*," and "*clap*," etc.) by combing both body motion and hand gesture information.

Another future research topic is to recognize human gestures from the egocentric perspectives of the robot. In the proposed GUHRI system, Kinect is employed as the vision sensor with fixed position. This system setup has some limits for real applications under challenging conditions. That is, the robot cannot change her viewpoint due to the fixed position of Kinect. In this case, the robot is not able to always acquire the optimal viewpoint to capture human body gesture information. Actually, due to this viewpoint reason, body occlusion may happen that will seriously confuse the accurate body joint position extraction. One feasible solution for this problem is to capture the human body gesture information from the robot's egocentric perspectives. In this way, the robot can change her viewpoint accordingly. However, to achieve good results, some new challenges need to be solved; one main problem is how to distinguish camera motion and real body motion.

In addition, how to integrate the verbal clues in the GURHI system to further enhance the human–robot interaction performance is also what we are concerned about in the future work. Making the robot "see" and "listen" will let her become more autonomous and humanlike.

2.6 Conclusion

The GUHRI system, a novel body gesture understanding and human–robot interaction system, is proposed in this paper. A set of 12 human upper body gestures with and without human–object interactions can be understood by the robot. Meanwhile, the robot can express herself by using a combination of body movements, facial expressions, and verbal language simultaneously, aiming to give the users a natural and vivid experience.

A new combination of sensors is proposed. That is, CyberGlove II and Kinect are combined to capture the head, arm, and hand posture simultaneously. An effective and real-time gesture recognition method is also proposed. In the experiment, a human upper body gesture dataset is built. The experimental results demonstrate the effectiveness and efficiency of our gesture recognition method.

So far, the gestures involved in GUHRI system have been static ones, e.g., "*have question*," "*praise*," "*call*," "*drink*," etc. As the future work, we plan to enable the robot to understand dynamic gestures such as "*wave hand*," "*say no*," "*clap*," etc. Speech recognition can be further added to make the interaction more natural.

References

1. Adamo-Villani N, Heisler J, Arns L (2007) Two gesture recognition systems for immersive math education of the deaf. In: Proceedings of the first international conference on immersive telecommunications (ICIT 2007). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), p 9
2. Beck A, Cañamero L, Hiolle A, Damiano L, Cosi P, Tesser F, Sommavilla G (2013) Interpretation of emotional body language displayed by a humanoid robot: a case study with children. Int J Soc Robot: 1–10
3. Belongies S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Int 24(4):509–522
4. Berman S, Stern H (2012) Sensors for gesture recognition systems. IEEE Trans Syst Man Cybern: Appl Rev 42(3):277–290
5. Brethes L, Menezes P, Lerasle F, Hayet J (2004) Face tracking and hand gesture recognition for human-robot interaction. In: Proceedings of IEEE conference on robotics and automation (ICRA 2004), IEEE, vol 2. pp 1901–1906
6. Cañamero L, Fredslund J (2001) I show you how i like you—can you read it in my face? [robotics]. IEEE Trans Syst Man Cybern: Syst Hum 31(5):454–459
7. Cassell J et al. (2000) Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents. Embodied conversational agents, pp 1–27
8. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR 2005), vol 1. pp 539–546
9. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
10. Dautenhahn K (2007) Socially intelligent robots: dimensions of human-robot interaction. Philos Trans Royal Soc B: Biol Sci 362(1480):679–704
11. Faber F, Bennewitz M, Eppner C, Gorog A, Gonsior C, Joho D, Schreiber M, Behnke S (2009) The humanoid museum tour guide robotinho. In: Proceedings of IEEE symposium on robot and human interactive communication (RO-MAN 2009), IEEE, pp 891–896
12. Fisher RA (1936) The use of multiple measures in taxonomic problems. Ann Eugenics 7:179–188
13. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3):143–166
14. Goodrich MA, Schultz AC (2007) Human-robot interaction: a survey. Found Trends Hum-Comput Interact 1(3):203–275
15. Immersion (2010) Cyberglove II specfications
16. Jolliffe IT (1986) Principal component analysis. Springer, London
17. Krämer NC, Tietz B, Bente G (2003) Effects of embodied interface agents and their gestural activity. In: Intelligent virtual agents. Springer, London, pp 292–300
18. Lu G, Shark L-K, Hall G, Zeshan U (2012) Immersive manipulation of virtual objects through glove-based hand gesture interaction. Virtual Reality 16(3):243–252
19. Mehrabian A (1971) Silent messages
20. Müller M, Röder T (2006) Motion templates for automatic classification and retrieval of motion capture data. In: Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on computer animation (SCA 2006), Eurographics Association, pp 137–146

21. Nickel K, Stiefelhagen R (2007) Visual recognition of pointing gestures for human-robot interaction. Image Vis Comput 25(12):1875–1884
22. Osborne JW, Costello AB (2004) Sample size and subject to item ratio in principal components analysis. Pract Assess, Res Eval 9(11):8
23. Perzanowski D, Schultz AC, Adams W, Marsh E, Bugajska M (2001) Building a multimodal human-robot interface. IEEE Intell Syst 16(1):16–21
24. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR 2011), pp 1297–1304
25. Smith LB, Breazeal C (2007) The dynamic lift of developmental process. Dev Sci 10(1):61–68
26. Spiliotopoulos D, Androutsopoulos I, Spyropoulos CD (2001) Human-robot interaction based on spoken natural language dialogue. In: Proceedings of the European workshop on service and humanoid robots, pp 25–27
27. Stiefelhagen R, Ekenel HK, Fugen C, Gieselmann P, Holzapfel H, Kraft F, Nickel K, Voit M, Waibel A (2007) Enabling multimodal human-robot interaction for the karlsruhe humanoid robot. IEEE Trans Robot 23(5):840–851
28. Stiefelhagen R, Fugen C, Gieselmann R, Holzapfel H, Nickel K, Waibel A (2004) Natural human-robot interaction using speech, head pose and gestures. In: Proceedings of IEEE conference on intelligent robots and systems (IROS 2004), IEEE, vol 3, pp 2422–2427
29. Teleb H, Chang G (2012) Data glove integration with 3d virtual environments. In: Proceeedings of international conference on systems and informatics (ICSAI 2012), IEEE, pp 107–112
30. Waldherr S, Romero R, Thrun S (2000) A gesture based interface for human-robot interaction. Auton Robots 9(2):151–173
31. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR 2012), pp 1290–1297
32. Wang J, Liu Z, Chorowski J, Chen Z, Wu Y (2012) Robust 3d action recognition with random occupancy patterns. In: Proceedings of European conference on computer vision (ECCV). Springer, London, pp 872–885
33. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244
34. Xiao Y, Cao Z, Zhuo W (2011) Type-2 fuzzy thresholding using glsc histogram of human visual nonlinearity characteristics. Opt. Express 19(11):10656–10672
35. Xiao Y, Cao Z, Yuan J (2014) Entropic image thresholding based on GLGM histogram. Pattern Recogn Lett 40:47–55
36. Xiao Y, Wu J, Yuan J (2014) mCENTRIST: a multi-channel feature generation mechanism for scene categorization. IEEE Trans Image Process 23(2):823–836
37. Xiao Y, Yuan J, Thalmann D (2013) Human-virtual human interaction by upper body gesture understanding. In: Proceedings of the 19th ACM symposium on virtual reality software and technology (VRST 2013), pp 133–142. ACM, Las Vegas
38. Xiao Y, Zhang Z, Beck A, Yuan J, Thalmann D (2014) Human-robot interaction by understanding upper body gestures. Presence: teleoperators and virtual environments (Accepted)

http://www.springer.com/978-3-319-19946-7

Context Aware Human-Robot and Human-Agent Interaction

Magnenat-Thalmann, N.; Yuan, J.; Thalmann, D.; You, B.-J. (Eds.)

2016, XIII, 298 p., Hardcover
ISBN: 978-3-319-19946-7