# A Survey on Perception Methods for Human–Robot Interaction in Social Robots

**Article** *in* International Journal of Social Robotics · January 2014

DOI: 10.1007/s12369-013-0199-6

Marcelo H. Ang Jr
National University of Singapore
**246** PUBLICATIONS **3,327** CITATIONS

Aun-Neow Poo
National University of Singapore
**155** PUBLICATIONS **3,500** CITATIONS

Int J Soc Robot (2014) 6:85–119
DOI 10.1007/s12369-013-0199-6

SURVEY

# A Survey on Perception Methods for Human–Robot Interaction in Social Robots

**Haibin Yan · Marcelo H. Ang Jr. · Aun Neow Poo**

Accepted: 30 May 2013 / Published online: 4 July 2013

**Abstract** For human–robot interaction (HRI), perception is one of the most important capabilities. This paper reviews several widely used perception methods of HRI in social robots. Specifically, we investigate general perception tasks crucial for HRI, such as where the objects are located in the rooms, what objects are in the scene, and how they interact with humans. We first enumerate representative social robots and summarize the most three important perception methods from these robots: feature extraction, dimensionality reduction, and semantic understanding. For feature extraction, four widely used signals including visual-based, audio-based, tactile-based and rang sensors-based are reviewed, and they are compared based on their advantages and disadvantages. For dimensionality reduction, representative methods including principle component analysis (PCA), linear discriminant analysis (LDA), and locality preserving projections (LPP) are reviewed. For semantic understanding, conventional techniques for several typical applications such as object recognition, object tracking, object segmentation, and speaker localization are discussed, and their characteristics and limitations are also analyzed. Moreover, several popular data sets used in social robotics and published semantic understanding results are analyzed and compared in light of our analysis of HRI perception methods. Lastly, we suggest important future work to analyze fundamental questions on perception methods in HRI.

**Keywords** Social robots · Human–robot interaction · Perception · Survey

H. Yan (✉) · M.H. Ang Jr. · A.N. Poo
Department of Mechanical Engineering, National University of Singapore, Singapore 117576, Singapore
e-mail: eyanhaibin@gmail.com

## 1 Introduction

Social robotics, an important branch of robotics, has recently drawn increasing attention in many disciplines, such as computer vision, artificial intelligence, and mechatronics, and has emerged as an interdisciplinary undertaking. While a number of social robots have been developed, a formal definition of social robot remains unclear and different practitioners have defined it from different perspectives. For example, Fong *et al.* [1] define that social robots are able to recognize each other and engage in social interactions; Breazeal *et al.* [2] explain that a social robot is a robot which is able to communicate with humans in a personal way; Bartneck and Forlizzi [3] describe that a social robot is an autonomous or semi-autonomous robot that interacts with humans by following some social behaviors; Hegel *et al.* [4] define that a social robot is a combination of a robot and a social interface. In Wikipedia, 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, a common characteristic can be reflected and we define in this paper a social robot as follows:

> *“A social robot is a robot which can execute designated tasks, and the necessary condition turning a robot into a social robot is the ability to interact with humans by adhering to certain social cues and rules.”*

It is generally believed that human–robot interaction (HRI) is the heart of a social robot, and the interaction capability is the most important factor for a social robot. There have been several attempts on HRI of social robots in recent years, and two representative examples are Kismet (developed by MIT in 2002) [6] and ASIMO (developed by the Honda company in 2003) [7]. Breazeal and colleagues in

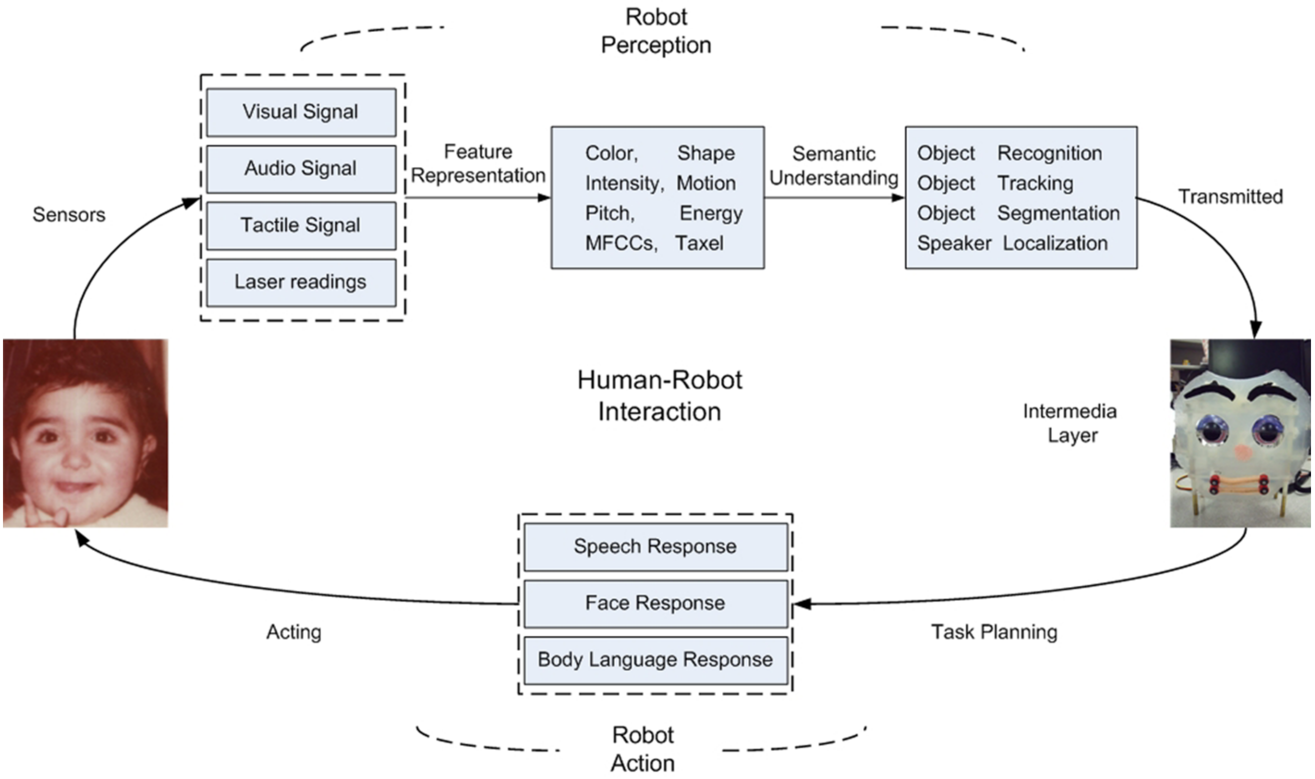

**Fig. 1** Flow chart of human–robot interaction. The above and below loops consist of the perception and action parts of the human–robot interaction system. First, the perception part captures the raw signals of the person made by using different sensors such as cameras, microphones, laser reading and so on. Then, some feature representation methods are used to characterize the signatures of the raw signals captured. Based on the extracted features, some semantic understanding tasks such as object tracking and recognition, speaker localization and recognition, and scene understanding are conducted. Having obtained the results of these semantic understanding tasks, the robot performs some tasks such as face and body response to interact with the subject

the MIT Media Lab developed Kismet, which has a human-like head to communicate with humans. Honda developed ASIMO which also demonstrated human-like characteristics to assist humans. More recently, a number of social robots have been designed and already or potentially applied to people's everyday lives as companions, assistants, and entertainment toys [8]. For example, RoboX [9] is a tour-guide robot at the Swiss National Exhibition Expo.02, Sony AIBO [10] has entertained humans to bring happiness, and Kismet has assisted people for social interaction studies. In this survey, we mainly focus on the interaction between a social robot and humans, also called human–robot interaction (HRI) in social robots.

Generally, HRI of a social robot consists of three parts: perception, action and an "intermediate" mechanism. In this paper, perception refers to an environmental information acquisition and analysis module; action means the responses a robot made after it receives motor-control signals; and the intermediate mechanism is equivalent to "a robot's brain" connecting perception and action to produce motor-control signals according to the results from the perception analysis module. Figure 1 shows the flow chart of a general HRI framework in a social robot (the face image of the child is from the public FG-NET Aging Database [11]).[1] We can see from this figure that the perceptual system dominates the whole HRI because it acts as a bridge between a social robot and the outside environment. Only the robot accurately understands the surrounding world, it can give meaningful responses to the correct subject. Due to high potentials and great importance of a perceptual system, we review in this paper state-of-the-art perception methods of HRI in social robots from three aspects: feature extraction, dimensionality reduction, and semantic understanding.[2]

[1] In this paper, we do not differentiate feature representation and feature extraction while there are some slightly differences. Generally speaking, feature representation methods focus on extracting low-level features and feature extraction methods aim at extracting middle- and high-level features.

[2] In some social robots, semantic understanding is included in the intermediate mechanism. In this paper, we consider it as a part of the perception system since it is indirectly fulfilled on the acquired signals and the understanding can represent the outside environments that a robot really wants to know.

Features refer to the information conveyed by raw signals, such as energy, color, and texture. Generally speaking, there are four classes of signals captured by a social robot: visual-based, audio-based, tactile-based, and range sensors-based. They are collected by cameras, microphones, tactile sensors, and laser range finders, respectively. Semantic understanding includes advanced analysis on these extracted features, such as sound localization, face detection, and emotion recognition.[3] In this paper, we review several typical semantic understanding examples such as object recognition, object tracking, object segmentation, and speaker localization, and discuss their characteristics and limitations.[4]

The contributions of this work are summarized and highlighted as follows:

1. We have described a survey of perception methods for HRI in social robots, which summarizes perception methods from three aspects: feature extraction, dimensionality reduction, and semantic understanding. Moreover, we have also presented several popular data sets and published results to compare the advantage and disadvantages of different existing methods.
2. We have reviewed some representative social robots to show how the perception methods work in existing social robots, which can also effectively show their importance for HRI.
3. We have pointed out the limitations of existing HRI systems and suggested some promising directions for future work in this area.

The paper is organized as follows. Section 2 reviews some representative social robots to show how the perception methods work in existing social robots. Section 3 summarizes state-of-the-art perception methods. Moreover, several popular data sets used in social robotics are analyzed and compared in light of our analysis of HRI perception methods, and a brief comparison of some published results is also presented. Section 4 concludes the survey and suggests challenges and future work.

## 2 Review of Representative Social Robots

While many challenges are encountered when applying social robots in real-world applications, there are still some social robots developed to assist our daily life, such as Kismet, iCub, and Robovie, as well as some commercially available social robots including Sony's AIBO and NEC's PaPeRo. In this section, we introduce some representative social robots and show how the perception methods work in these robots. Table 1 compares their basic functions and characteristics (the sensor and the core algorithm).[5] These methods enumerated in this table can work in real-time and are fully autonomous.

### 2.1 Robots as Test Subjects

#### *2.1.1 Social Development*

For the research of social development, infants and young children are the popular studied subjects. Hence, social robots as test subjects are generally designed as infant-like or child-like, which have some learning skills such as imitation and joint attention.

Cog [12] is a representative robot, which is developed for human cognition and developmental psychology. It has 22 DOFs distributed on the arms, torso, neck, and eyes. This robot can manipulate objects, show head postures, and move eyes. To achieve these goals, visual signals were adopted for face detection and object segmentation [12, 13].

iCub is another representative social robot, and its objective is to offer a platform for cognition investigation. It comes from RoboCub, a 5-year project funded by the European Commission through Unit E5 "Cognitive Systems, Interaction & Robotics" [14]. iCub is particularly designed as a 3–4 years old child. It can interact with the outside environments with the head, neck, arms, torso, and legs. For example, iCub can follow objects by orientating its head and eyes, dexterously manipulate objects with its hands, crawl and sit up like a child [15]. Human detection, object recognition, and sound localization were implemented in this robot, respectively [16–18, 226].

#### *2.1.2 Social Interaction*

An important way to learn new knowledge from environments is social interactions. To investigate social interactions between humans, a natural approach is to design human-like robots. According to the pre-defined personality, robots make reasonable responses through speech or body languages when interacting with humans.

Kismet [19], a well-known robot head, is designed to investigate social interactions between caregivers and babies. It interacts with humans through facial expression, body gesture, and vocal babbles. Its perception abilities include

[3] In some areas, emotion recognition is also called affective computing and emotion is referred to as affective state. In this paper, we still use emotion recognition because it is more commonly used in the literature.

[4] In this survey, we mainly focus on four widely used signals used in existing social robots. Hence, other types of signals such as GPS, temperature and pain of humans will not be discussed.

[5] Please note that the core algorithm refers to the approach used for some semantic understanding tasks such as detection/classification/prediction in HRI by using the extracted features.

**Table 1** Taxonomy of perception methods on existing representative social robots

| Social robot | Sensor | Feature | Semantic understanding task | Core algorithm |
|---|---|---|---|---|
| Kismet | camera | motion, color [20, 21] | object detection/recognition | nearest neighbor |
| | microphone | pitch, energy [22] | emotion recognition | Gaussian mixture model |
| Cog | camera | intensity [13] | face detection | template matching |
| | camera | motion [12] | object segmentation | correlation analysis |
| iCub | camera | intensity, shape [16] | human and object detection | mutual information |
| | camera | SIFT-based [17] | object recognition | hough transform + least-mean-squares |
| | microphone | ITD, ILD, notches [18] | sound localization | mapping |
| GRACE | laser reading | shape [27] | people in line perception | nearest neighbor |
| | camera | color, shape [28] | human tracking | Kalman filter |
| | microphone | IBM's ViaVoice [27] | speech recognition | IBM's ViaVoice |
| Robox | laser reading | raw data [30] | motion detection | nearest neighbor |
| | camera | shape [30] | object tracking | Kalman filter |
| | camera | color [31] | face detection | Heuristic filter |
| | microphone | Viterbi-based [9] | speech recognition | hidden Markov model |
| Reckham | camera | Harr-like [32] | face detection | boosting |
| | camera | learning-based [33] | face recognition | nearest neighbor |
| | camera | color, shape, motion [33] | human tracking | condensation |
| | camera | color, shape [33] | gesture classification | condensation |
| Robovie | camera | color, learning-based [34] | gesture recognition | nearest neighbor |
| | laser reading | human trajectories [35] | human behavior analysis | support vector machine + $k$-means |
| | laser reading | raw data [35] | human tracking | clustering and particle filter |
| RUBI | microphone | STBF-based [37] | emotion recognition | boosting |
| | camera | Gabor-based [38] | facial expression recognition | boosting + support vector machine |
| ARMARIII | camera | DCT-based [42] | face verification | nearest neighbor |
| | camera | intensity [43] | speaker tracking | particle filter |
| | microphone | time delay [43] | speaker tracking | particle filter |
| | microphone | learning-based [44] | sound classification | hidden Markov model |
| | camera | intensity, shape [45] | head pose estimation | Neural Network |
| | camera | color, intensity [46] | human gesture recognition | Neural Network + hidden Markov model |
| PaPero | camera | shape, 3D model [48] | face/eye detection, recognition | template match |
| | microphone | fiter banks [49] | speech recognition | hidden Markov model |
| Huggable | tactile | calculation-based [50] | touch classification | Neural Network |
| MEXI | camera | color, intensity [13, 52] | face detection | template match |
| | microphone | frequency, energy [53, 54] | emotion recognition | fuzzy logic |
| ROMAN | microphone | raw data [56] | sound localization | beam-forming |
| | microphone | energy [56] | object tracking | particle filter |
| | camera | feature point [57] | emotion recognition | facial action coding system |
| BARTHOC | microphone | pitch, energy [59] | emotion recognition | bayes |
| | microphone | temporal shift [60] | voice detection | cross-power spectrum phase |
| | camera | haar-like [60] | human tracking | boosting |
| BIRON | laser reading | raw data [62] | human tracking | nearest neighbor |
| | camera | color, haar-like [62, 63] | human tracking | boosting |
| | microphone | time delay [63] | sound localization | cross-power spectrum phase |
| Fritz | camera | haar-like [64] | face detection and tracking | boosting + Kalman filter |
| | microphone | time delay [64] | speaker localization | cross-power spectrum phase |

object detection, object recognition, and emotion recognition [20–22]. Kismet has been a popular prototype in many research groups to develop their social robots [23, 24].

Besides human-like appearances, MIT Media Lab in collaboration with Stan Winston Studio and DARPA has developed a robot named Leonardo with animatronic characteristics [25]. Leonardo has a highly mobile face and body, and can learn new skills and cooperate with users by social interactions. It is able to remember the users, execute the verbal commands, and share attentions. There are several perception abilities on this robot, such as face recognition, speech recognition, object recognition and tracking, and tactile perception.

## 2.2 Socially Assistive Robotics

There are some socially assistive robots which provide services to humans in public places or domestic environments [26]. Since it is closely related to people's lives, more recognition abilities and social cues are required in these robots. Moreover, due to the difference between the public places and the domestic environments, the perception tasks are also distinct.

### *2.2.1 Robots in Public Places*

Robots in pubic places refer to the robots used in museums, supermarkets, shopping centers, and childhood education centers. Different places require different tasks. For example, if the robot is applied in museums, it is used to introduce the exhibits to humans; when it is applied for childhood education, it is used to help teacher to organize or teach children.

GRACE is a representative robot which is used in public places. It firstly appeared in AAAI 2002 to perform the Robot Challenge, where the tasks were to navigate the registration desk of the conference center, come to the conference room, and give a presentation [27]. In AAAI 2005, GRACE's task was extended to find a person [28]. In both AAAI 2002 and 2005, GRACE successfully completed speech recognition and human tracking in real-world environments.

RoboX, a tour-guide robot, appeared at the Swiss National Exhibition *Expo.02* [9], and Rackham, another tour-guide robot, appeared in Mission BioSpace exhibition [29]. Their task is to present exhibits for tourists. To implement such task, they need to navigate environments, introduce themselves, ask visitors to choose a visiting destination, and give the corresponding presentations. The perception tasks include motion detection and tracking, face detection and recognition, speech recognition, and gesture classification [30–33].

Robovie is a robot with different functions. Robovie II aims to help the elders shop in the Apita-Seikadai supermarket in Kyoto, and Robovie III aims to provide directions for people and invite customers to visit the shop. Robovie II includes face and gesture recognition to recognize human faces and hand poses of the users. Robovie III contains human behavior analysis to identify the users' walking styles, and directions [34, 35].

RUBI is a three-feet tall robot. It consists of a head, two arms and 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, interacting 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 [36]. It contains some perception functions such as face detection and tracking, and emotion recognition [37, 38].

### *2.2.2 Robots in Domestic Environments*

Robots in domestic environments mean the robots used at home. Christensen [39] has summarized three domestic environments: entertainment, everyday tasks, and assistance to the elder and handicapped. Lohse and colleagues [40] have proposed a functionality-based categorization including health care, companionship, entertainment, toy, pet, and personal assistants. Different from the robots which are used in various and different public places, the robots in domestic environments are generally employed at home. Hence, their users are pre-specified.

ARMAR III, developed in University of Karlsruhe, can execute tasks in household environments [41]. The robot can open and close a dishwasher, pick up cups and dishwares, place them anywhere within reach, and plug an electrical appliance into the wall in a kitchen [42–46]. Moreover, it can interact with humans by using speech and gesture recognition. To achieve these goals, ARMAR III has a head, two arms, two hands, a torso and a mobile platform, to perform human verification and tracking, head pose estimation, and sound classification.

PaPeRo is a social robot designed by the NEC Corporation and has been commercially available. It is a personal robot which can care for children and provide assistance to elders. Several applications were developed: speech conversation, face recognition, touching reaction, roll-call and quiz game designing, communications through phone or PC, learning greetings, and story telling [47–49]. Moreover, speakers and LEDs are used to produce speech and songs and display PaPeRo's internal status, respectively.

Huggable [50] was developed by the personal robots research group from the MIT media lab. Different from the above described robots, it uses tactile-based signals for healthcare, education, and social communication applications. 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 to sense the surrounding environment. 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 elder and children through a web interface.

### 2.3 Robots for Studying Human–Robot Interaction

HRI plays an important role on the above robots even if they are developed for different applications. HRI can help a robot to learn new knowledge, obtain useful information, and attract persons. Motivated by these significant roles, more and more researchers have focused on HRI in social robots and developed several social robots, such as MEXI, ROMAN, BARTHOC, and Fritz for HRI studies.

MEXI [51] is a robot head to detect faces and recognize emotions. It responds to faces and emotions with different facial expressions such as smiling, sulking, and looking around [52–54]. Additionally, it uses a commercially available software to respond with speech. Similar to MEXI, ROMAN [55] is another humanoid robot head. ROMAN can follow the detected human [56] using its eyes and neck, and non-verbally communicate with humans using facial expressions [57].

BARTHOC and BARTHOC Junior (Jr) are two upper-body humanoid robots for studying HRI [58]. The main differences between them include the sizes and the weights. BARTHOC mimics adult person, and BARTHOC Jr imitates a four-year old child. Both of them can perform emotion recognition, voice detection and human tracking [59, 60]. BIRON is another robot developed by Bielefeld University [61], and has ported its several basic functions to BARTHOC such as human tracking [62, 63]. Since the appearances of BARTHOC and BARTHOC Jr are more human-like than that of BIRON, speech-based emotion recognition has been developed and applied in BARTHOC Jr. With the recognized results, the robot can mirror human's affective states by its own facial images.

Fritz [64] is another body-based social robot developed for HRI. It was originally developed to play soccer, and is now a platform to study multimodal communication with humans. It has perception abilities such as face detection, face tracking, and speaker localization. By using body gestures, facial expressions and synthesized speech, the robot can attract person's interest on communication.

Besides the above mentioned robots, there are several other social robots which can also implement semantic understanding task, such as Cherry and Petra [65–67], CERO [68, 69], Keepon [70], Paro [71], Probo [72–74], ASIMO [7], iCat [75, 76], AIBO [10, 77], Albert Einstein [78], WE-4RII [79–81], and Maggie [82, 83]. Moreover, some of them have been commercialized. Table 2 lists some properties of these social robots, in which the sensors, semantic understanding tasks, appearances, and functions are presented.

To allow robot researchers to focus on the specific fields, some companies and research institutions have developed the robot platforms from both hardware and software aspects. These robot platforms can implement several basic functions. For example, HOAP series from Fujitsu [84], QRIO from Sony [85], HUBO (KHR-3) from KAIST [86], and HRP series from Kawada Industries [87] are humanoid robot platforms. While they are of different appearances and sizes, their basic functions are similar. These robots have perception abilities like object recognition and tracking, sound localization and recognition. QRIO can also recognize the users' voice and face. Another two robot platforms are PR2 from Willow Garage [88] and YouBot from KUDA [89]. They have been commercially available now. PR2 is designed as a personal robot with mobility. It can navigate human environments and manipulate some objects in the environments. For YouBot, it consists of a manipulator and a mobile base. Hence, it has mobility and can grasp some objects. Based on these developed robot platforms, several perception tasks such as emotion recognition and gesture recognition can be included for HRI study.

### 2.4 Discussion

In this section, we reviewed several representative social robots according to their different applications including robots as test subjects, robots providing social assistance, and robots for studying HRI. We mainly focused on the employed perception methods on HRI of these robots. 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". A significant challenge of the study is how to achieve natural communication between humans and robots. To address this, one primary and key factor is using different perception methods in HRI.

As we described above, different social robot applications affect the design of perception systems. For the robots as test subjects and for studying HRI, they are usually used in lab environments which are much simpler than the real-world environments, and the perception tasks mainly include human detection and tracking, face detection, recognition, and tracking, gesture recognition, sound localization and recognition, and emotion recognition. In addition, when humans use objects such as toys to interact with the robots,

**Table 2** Properties of some other social robots

| Application | Social robot | Sensor | Semantic understanding task | Appearance | Function | Commercialized |
|---|---|---|---|---|---|---|
| Public service | Cherry and Petra | camera, touch screen | face recognition, user interface | mobile base (AmigoBot and PeopleBot) | to fetch, carry, and guide in offices | No |
| | CERO | microphone, PDA | speech interface, mobile, networked PDA-based interface | mobile base (Nomadic Super Scout platform) | to fetch and carry in offices | No |
| | Keepon | camera, microphone | face and colorful toys detection and object tracking, rhythms recognition | a small yellow snowman with a black cylinder | to study social development in research institutes, autism therapy in care centers, and play with children in a playroom | Yes |
| | Paro | microphone, tactile, temperature, posture | sound localization, speech recognition, touch and elevation detection | a baby harp seal with pure white fur | to help the elderly or children's therapy in hospitals or nursing homes | Yes |
| | Probo | camera, microphone, sensitive skin | face and object detection, sound localization and identification, touching detection and recognition | a imaginary green elephant with a touch screen | to communicate with children in hospitals | No |
| | ASIMO | camera, microphone, tactile | face recognition, human tracking, sound discrimination, touch sensing | whole-body humanoid robot | to provide service in public places | No |
| Domestic service | iCat | camera, microphone, tactile | face recognition, head tracking, sound localization, speech recognition, touch sensing | a cartoon cat without mobile ability | to be a family companion to control in-home devices | Yes |
| | AIBO | camera, microphone, tactile | face and objects detection and recognition, speech recognition, user's voice detection and recognition, touch sensing | a robot dog | to entertain users and used as a research platform | Yes |
| HRI | Albert Einstein | camera, microphone | face detection, tracking and identification, object tracking, behavior and facial expression recognition, speech recognition | robot head | to study HRI | No |
| | WE-4RII | camera, microphone, force sensor | colorful object detection and tracking, sound localization, touching detection and recognition | upper-body humanoid robot | to study HRI | No |
| | Maggie | camera, microphone, tactile sensor | human detection and tracking, face recognition, speech processing, touching sensing | whole-body humanoid robot | to study HRI | No |
| Robot platform | PR2 | camera, laser | object resonation | whole-body humanoid robot | to be a robotic platform | Yes |
| | YouBot | camera, range sensor | object resonation | mobile base with a manipulator | to be a robotic platform | Yes |

the robots should be able to detect and track the objects. If humans want to have physical contacts with the robots, the robots should also have the ability of tactile detection and classification. For these perception tasks, visual signals acquired by cameras, audio signals acquired by microphones, and tactile signals acquired by tactile sensors are usually employed. Since the number of users under such scenarios is limited, the environment is comparatively simple, and the robots normally have no mobility, the extracted features and semantic understanding methods are usually simple.

For socially assistive robotics, since they could be used in home, offices, or hospitals, the extracted features and semantic understanding methods are normally different from those used in robots as test subjects and for studying HRI. For example, if the robot is applied in a shopping center, its human tracking system usually requires to simultaneously track several persons rather than only one user. Usually, due to the complex application environments like uncontrolled illumination and background, it will need more robust features and semantic understanding methods to help the robot successfully fulfill the tasks. In addition, because of the mobility of several socially assistive robotics, besides visual, audio, and tactile signals, laser reading acquired by range finders is another useful modality to interact with the outside environment for the robots.

Generally speaking, the selection of sensors and feature extraction methods depend on the semantic understanding tasks, and semantic understanding tasks depend heavily on the application scenarios of robots. For instance, the semantic understanding tasks for the robot used in a museum and that used in a childhood education center are different due to different objectives. Correspondingly, the sensors and feature extraction methods are also different for distinct semantic understanding tasks. Four basic rules are applied to select the sensors and extract features in social robots: (1) high stability, (2) fast speed, (3) high accuracy, and (4) high autonomy.

## 3 Perception Methods in HRI

State-of-the-art perception methods of social robots can be mainly classified into three steps: feature extraction, dimensionality reduction, and semantic understanding. The aim of feature extraction is to convert the raw signals from sensors to feature descriptors for subsequent understanding tasks. Four widely used signals including visual-based, audio-based, tactile-based, and range sensors-based are employed for feature extraction. The aim of dimensionality reduction is to reduce the complexity of computation after feature extraction. The aim of semantic understanding is to infer the objects or human behaviors from the extracted features. Typical semantic understanding tasks include object detection and recognition, human tracking and identification, speech recognition, emotion recognition, and touching detection and recognition.[6] We present some representative feature extraction and dimensionality reduction methods and semantic understanding examples in this section. The main abbreviations used in this section are listed in Table 3.

**Table 3** Main abbreviations used in the paper

| Abbreviation | Full name |
|---|---|
| LBP | local binary pattern |
| SIFT | scale invariant feature transform |
| AHT | adaptive hough transform |
| EOH | edge orientation histograms |
| EMD | elementary motion detector |
| PCA | principle component analysis |
| LDA | local discriminant analysis |
| LPP | locality preserving projections |
| DCT | discrete cosine transform |
| MFCCs | mel frequency cepstral coefficients |
| ITD | interaural time difference |
| IID | interaural intensity difference |
| ISD | interaural spectral difference |
| ICA | independent component analysis |
| HMM | hidden Markov model |
| GMMs | Gaussian mixture models |

### 3.1 Feature Extraction

As shown in Table 1, four kinds of sensors including cameras, microphones, tactile sensors, and laser range finders are widely used for signal acquisition in HRI. Hence, we review state-of-the-art feature extraction methods based on these four categories.

#### *3.1.1 Visual-Based Methods*

Visual signals have been widely used in social robots to achieve semantic understanding tasks such as face detection and recognition, human tracking and identification, facial expression recognition, and gesture classification. The reason why visual signals are popular in social robots is that most information (∼75 %) received everyday for human beings are visual signals [91]. Motivated by this fact, most social robots use visual signals to conduct human-like perception. According to the camera types, visual signals consist of 2D-based and 3D-based. 3D visual signals are usually collected by Kinect [92] or stereo cameras such as the Bumblebee [93], and have become an important modality in HRI perception recently.

*1 2D-based*

A large number of 2D visual-based feature extraction methods have been proposed in social robots. Some of these visual features could reflect characteristics of objects such as color, shape, and texture of the raw signals acquired, which are closely related to objects' attributes and could be named attribute-based features. Some of these features are designed

[6] Please note that semantic understanding presented in this paper refers to the interaction between humans and robots, and hence other tasks such as navigation in mobile robots and interaction between robots and the environments are not included in this paper.

based on algorithms which could be called algorithm-based features such as filtered feature and Haar-like feature. Different applications usually require different visual features. For example, color is effective for object detection because objects to be detected usually demonstrate different color distributions; shape is useful for human detection and tracking as the shape information of humans and other objects are exclusive; texture plays an important role in face recognition since human faces have a specific texture structure compared with other objects.

*A. Color*

Color is an important feature to represent objects, and two typical application examples include color-based object detection and color-based face detection. Since different color demonstrates different characteristics, the key to extract color feature is how to effectively employ the information of RGB (red, green, and blue) channels to better describe different colors. In Kismet's active vision system [20], it used some specified color transformation functions on RGB channels to extract color saliency features [94]. In MEXI's vision processing system [51], it applied region growing and merging method to obtain similar color regions for segmentation of objects of interest. While color is an effective feature to represent objects, it is easily affected by changing illumination. To address this problem, a large number of methods have been proposed. Jensen and colleagues [9] normalized the green and blue channels with the red channel, and then detected skin-color regions by selecting certain ranges of the color features in different color channels. Darrell *et al.* [95] applied a log color-opponent space to localize skin-color regions by using a Gaussian probability model, which is robust to shade from illumination. Wang *et al.* [96] employed a YCbCr color space to uncorrelate the intensity and chrominance, such that skin color is clustered into a small area even under a poor lighting condition. These examples have clearly shown the efficacy of color for feature extraction in social robots.

*B. Intensity*

In some applications, color may not be stable due to poor acquisition conditions. Intensity is another popular visual feature and has also been widely used in many social robots. Since color cameras are widely used nowadays, it is necessary to design an appropriate color transformation space to convert color signals to gray-scale ones, and the gray-scale value of each pixel in an image is referred as an intensity feature. There are a number of color transformation spaces and some of them have been successfully applied to social robots [20, 34, 94–96]. Please refer to [97, 98] for details.

Having obtained the gray-scale signals, intensity can be employed for feature representation. Ruesch and colleagues [99] utilized Mexican hat wavelets with different sizes in gray-scale images to extract conspicuous regions. Chen and Tiddeman [100] exploited horizontal and vertical integral projections to detect human's eyes by using the intensity information since the changing of the pixel intensity around the eyes is clearer than that of other regions in faces. Disparity information can also be obtained by using the intensity features, and has been used for object segmentation in some social robots [95, 101].

*C. Visual Texture*

Visual texture is an important property for visual signals, and different objects usually demonstrate different texture characteristics. A number of texture representation methods have been proposed for texture analysis and classification, and we believe that most of them can be extended to feature extraction in social robots. In this subsection, we only present several representative ones.

Song *et al.* [102] presented an image ratio feature representation method for facial expression recognition, which is robust to various illuminations. Given a point $p$ on a patch $j$, the expression ratio at this point is defined as

$$\Re = \frac{I'(u,v)_p(j)}{I(u,v)_p(j)} \tag{1}$$

where $(u,v)$ is the image coordinate of $p$, $I'(\cdot)$ and $I(\cdot)$ denote the intensity of the expressional and neutral images, respectively.

Local binary pattern (LBP) is another popular texture descriptor for feature representation. Inspired by Wang and He's work [103], Ojala *et al.* [104] proposed a novel LBP feature representation method for face recognition. The basic idea is as follows: for each pixel, its 8-neighborhood pixels are thresholded into 1 or 0 by comparing them with the center pixel. Then the binary sequence of the 8-neighborhoods is transferred into a decimal number (bit pattern states with upper left corner moving clockwise around center pixel), and the histogram with 256 bins of the processed image is used as the texture descriptor. To capture the dominant features, Ojala *et al.* [105] extended LBP to a parametric form $(P,R)$, which indicates that $P$ gray-scale values are equally distributed in a circle with radius $R$ to form circularly symmetric neighbor sets. To better characterize the property of texture information, uniform pattern is defined according to the number of spatial transitions that are bitwise $0/1$ changes in the calculated binary values. If there are at most two bitwise transitions from 0 to 1 or vice versa, the binary value is called a uniform pattern.

Recently, Jin *et al.* [106] developed an improved LBP (ILBP) method. Different from LBP, ILBP makes better use of the central pixel in the original LBP, and takes the mean of all gray values of elements as the threshold value. For the newly generated binary value of the central pixel, it was added to the most left position of the original binary string. The corresponding decimal value range would be changed from $[0,255]$ to $[0,510]$ for a $3 \times 3$ operator. Figure 2 shows the basic ideas of LBP and ILBP. Since LBP and ILBP are

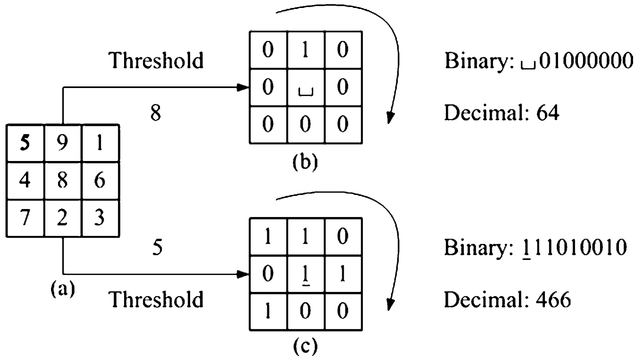

**Fig. 2** The LBP and ILBP operators where (**a**) original image patch (**b**) results of LBP operator (**c**) results of ILBP operators

robust to illumination changes, they have been widely used in many semantic understanding tasks such as face recognition and facial expression recognition [107–109].

More recently, Zhao and Pietikäinen [110] presented a spatiotemporal LBP that extends the original LBP to three orthogonal planes including $XY$, $XT$ and $YT$ for facial expression recognition, where $X$ and $Y$ are the width and the height of each face image, and $T$ is the length of the image sequence. Then, the histogram equation was modified to

$$H_{b,c,j,i} = \sum_{x,y,t} I\big(f_j(x,y,t)=i\big),$$
$$i=0,\ldots,n_j-1;\ j=0,1,2, \tag{2}$$

where $n_j$ is the number of different labels produced by LBP in the $j$th plane, $b$ and $c$ are the indexes of rows and columns of image pixels.

Similar to LBP and ILBP, census transform (CT) and modified census transform (MCT) were also proposed for texture representation. The only difference between MCT and ILBP is that their orders of the generated binary strings are different. More details can be found in [96, 111, 112].

In addition to LBP and its extensions, Scale Invariant Feature Transform (SIFT) feature is another important local feature for texture description, and has been widely used in object recognition, robotic navigation, video tracking, and image matching [113]. The key idea of this method is to transform an image into a large collection of local feature descriptors that densely cover the image over the full range of scales and locations, such that it can detect local feature points that are invariant to translation, scaling, rotation, and small distortions [114].

There are four main steps in the SIFT method: scale-space extrema detection, keypoint localization, orientation assignment, and keypoint description. Scale-space extrema detection is to identify potential interest points that are invariant to varying scales and orientations. A difference-of-Gaussian function is generally applied, and the local extrema of function at different scales are potential interest points. Then, for each keypoint candidate, its accurate location is determined by using interpolation of nearby data, and the one with low contrast or poorly localized along an edge is removed. The remaining points are selected as keypoints. To obtain robust descriptors for local affine distortion, each keypoint is assigned one or more orientations based on a gradient orientation histogram computed in the neighborhood of the keypoint. Lastly, the keypoint descriptor can be obtained by using the information of local image gradients that are measured at the selected scale in the region around each keypoint. For the details, please refer to [113].

*D. Shape*

Shape is another important feature for visual signal representation, especially for facial image analysis and human detection. For example, Tian *et al.* [115] extracted facial components including lip, eyes, brow, cheek, and furrow as human face features. A representative shape descriptor is the well-known snake model, which has been widely used in many computer vision applications. The snake model is an energy-minimizing spline, and the energy is represented by [116]

$$\begin{aligned} E^*_{snake} &= \int_0^1 E_{snake}\big(v(s)\big)ds \\ &= \int_0^1 E_{int}\big(\big(v(s)\big) + E_{image}\big(v(s)\big) + E_{con}\big(v(s)\big)ds \end{aligned} \tag{3}$$

where $v(s) = (x(s), y(s))$ is the position of a snake, $E_{int}$ is the internal energy of the spline due to bending, $E_{image}$ represents the image forces, and $E_{con}$ gives rise to the external constraint forces. With the cooperation of the forces represented in energy function, a snake can capture features like lines and edges. After initializing snakes with facial features, they lock onto the capture region surrounding features and localize them accurately.

Adaptive hough transform (AHT) is another shape descriptor to characterize human cheeks and chins. Compared with the original hough transform (HT), AHT achieves better performance in terms of both storage and computational requirements. AHT exploits the idea of a flexible iterative "coarse to fine" accumulation and search strategy based on small accumulator array. In Illingworth and Kittler's work, identifying line and circular segments in images were described [117].

Recently, edge orientation histograms (EOH) has been a popular feature representation method in object recognition and face detection. For example, Levi and Weiss presented three EOH features [118]. The first one was defined as

$$A_{k1,k2}(R) = \frac{E_{k1}(R)+\epsilon}{E_{k2}(R)+\epsilon} \tag{4}$$

**Fig. 3** Sobel convolution kernels where (**a**) vertical direction (**b**) horizontal direction

| | | |
|---|---|---|
| -1 | 0 | +1 |
| -2 | 0 | +2 |
| -1 | 0 | +1 |

(a)

| | | |
|---|---|---|
| +1 | +2 | +1 |
| 0 | 0 | 0 |
| -1 | -2 | -1 |

(b)

where $E_k(R) = \sum_{(x,y)\in R} \psi_k(x, y)$, $R$ is a subwindow in an image, $\psi_k(x, y)$ is the value of the $k$th bin based on a preprocessed image by the Sobel operator [118], $\epsilon$ is a smoothing term, in which $R$'s orientation was captured. The second one represented dominant orientation features that were computed by $Bk(R) = \frac{E_k(R)+\epsilon}{\sum_i E_i(R)+\epsilon}$. When there is a dominant edge orientation, features $Bk(R)$ are better than $A_{k1,k2}(R)$. The third one was the symmetry feature to capture symmetry of an image, and it was defined as

$$Symm(R_1, R_2) = \frac{\sum_{k\in K} |E_k(R_1) - E_k(R_2)|}{size\ of(R_1)} \tag{5}$$

where $R_1$ and $R_2$ are rectangles of the same size and opposite positions of a symmetry axes.

Edge is a useful technique to describe the shape information of objects, and a number of edge detection operators have been proposed for shape representation. Here, we present one representative method: Sobel operator. By using the Sobel operator on gray-scale images, the magnitude response in horizontal and vertical derivations can be captured. The detailed procedure is to implement convolution between a 3 × 3 kernel and the original image in both the horizontal and vertical directions. The used kernels are shown in Fig. 3 [119]. Sobel operator has been applied in normalized head images to extract features for head pose estimation [45].

*E. Motion*

Motion features have been widely used for object detection, tracking and recognition [127–130]. Optical flow is a typical motion feature, which is the distribution of velocities of brightness patterns' movement in an image [120]. Many optical flow methods have been proposed for motion representation. Brox *et al.* [121] presented an method to compute the optical flow by minimizing an energy function expressed as $E(u, v) = E_{Data}(u, v) + \alpha E_{Smooth}(u, v)$, where the two terms are related to the appearance and smoothness of the image, respectively, and $\alpha$ controls the regularization term. Bab-Hadiashar and Suter [122] proposed two robust optical flow methods by using the Least Median Squares and Least Median Squares Orthogonal Distances, to detect the outliers and solve the inlier group, respectively.

In addition to optical flow, Kismet [20] used a threshold function to compute the absolute difference between continuous frames in a video sequences, and the positions with high intensity values in a binary map were represented as motion features. iCub explored motion information in its attention system [99], where motion was detected by using a biologically plausible model called elementary motion detector (EMD) [123]. There are five parts in EMD which are photoreceptors, high-pass filters, low-pass filters, multipliers, and subtractions, where high-pass and low-pass filters are used to remove constant illumination information from input visual signals and delay signals, and multipliers are used to correlate non-delayed and delayed signals from high-pass and low-pass filters.

There are several surveys of motion feature extraction methods [124–126], where Cedras and Shah [124] reviewed the development of motion detection and recognition in computer vision, Moeslund and Granum [125] investigated human motion detection, and Wang and Singh [126] surveyed motion analysis from the video processing aspect. Without loss of generalization, we believe most of these methods can be generalized and applied to social robots.

*F. Filtered Feature*

There are some filtering features used in visual feature extraction, such as discrete cosine transform (DCT) and Gabor wavelet. DCT decorrelates the original data by using the following transformation [146]

$$F(u, v) = \frac{1}{\sqrt{MN}}\alpha(u)\alpha(v) \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x, y) \times \cos\left(\frac{(2x+1)u\pi}{2M}\right) \times \cos\left(\frac{(2y+1)v\pi}{2N}\right) \tag{6}$$

where $\alpha(\omega) = \frac{1}{\sqrt{2}}$ when $\omega = 0$ and 1 otherwise, $u = 0, 1, \ldots, M$, $v = 0, 1, \ldots, N$, $f(x, y)$ is the image intensity function, and $M \times N$ is the size of the image.

The key point of DCT is the selection of DCT coefficients because these coefficients have distinct discriminative power. Particularly, there are three parts of DCT coefficients: low frequencies containing illumination information, middle frequencies reflecting image content information, and high frequencies representing noise and image details. In Dabbaghchian and colleagues' work, they applied a data-dependent approach to search discriminative coefficients for feature representation [146]. DCT features have been used in local feature-based face recognition [147].

Gabor wavelet is a popular feature extraction method for visual signal representation, and discriminative information is extracted by convoluting the original image with a set of Gabor kernels with different scales and orientations. A 2-D Gabor wavelet kernel is the product of an elliptical Gaussian envelope and a complex plane wave, defined as [148, 149]:

$$\psi_{\mu,v}(z) = \frac{\|k_{\mu,v}\|^2}{\sigma^2} e^{-\frac{\|k_{\mu,v}\|^2\|z\|^2}{2\sigma^2}} \left[e^{ik_{\mu,v}z} - e^{-\frac{\sigma^2}{2}}\right] \tag{7}$$

where $\mu$ and $v$ define the orientation and scale of the Gabor kernels, $z = z(x, y)$ is the variable in a complex spatial domain, $\|\cdot\|$ denotes the norm operator, and the wave vector

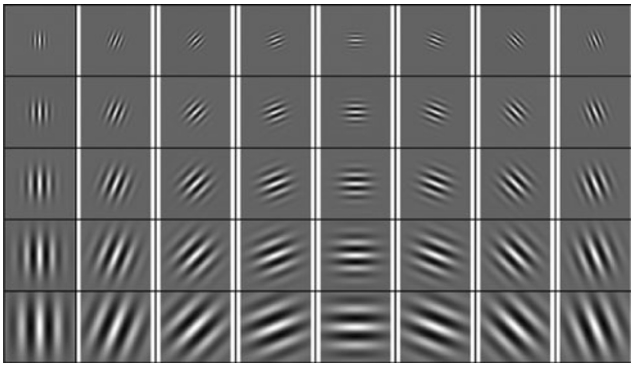

**Fig. 4** Real part of Gabor kernels at eight orientations and five scales [148]

$k_{\mu,\nu}$ is defined as follows:

$$k_{\mu,\nu} = k_\nu e^{j\phi_\mu} \tag{8}$$

where $k_\nu = k_{max}/f^\nu$, $\phi_\mu = \pi\mu/8$, $k_{max}$ is the maximum frequency, $f$ is the spacing factor between kernels in the frequency domain, and $\sigma$ is the standard deviation of Gaussian envelope determining the number of oscillations. For a given image $I$, the convolution of $I$ and a Gabor kernel $\psi_{\mu,\nu}$ is defined as

$$O_{\mu,\nu}(z) = I(z) * \psi_{\mu,\nu}(z) \tag{9}$$

where $O_{\mu,\nu}(z)$ is the convolution result corresponding to the Gabor kernel at orientation $\mu$ and scale $\nu$. Figure 4 shows the real part of the Gabor kernels.

Gabor-based features have shown good performance in face recognition and facial expression recognition [150, 151]. Usually, five spatial frequencies and eight orientations are used, and there will be a total of 40 Gabor kernel functions employed on each pixel of an image. Its computational cost is generally expensive. Moreover, only the magnitudes of Gabor wavelet coefficients are used as features because the phase information are sensitive to inaccurate alignment. Besides face-related applications, Gabor-based features have been used to detect directional saliency features in iCub's attention system [99].

*G. Haar-Like Feature*

Haar-like features are originally defined as the difference of the sum of pixels in two rectangular areas. One of the pioneering studies on Haar-like features is Viola and Jones' work [32]. In their work, three kinds of features were used for Haar-like feature extraction, including two-rectangle, three-rectangle, and four-rectangle features, as shown in Fig. 5. To speed up the computation, the integral image was presented to be an intermediatete representation of an image to conveniently provide the sum of the pixels.

Besides the conventional Haar-like features, Pavani *et al.* [152] proposed an extension of Haar-like features which is optimally weighted rectangle and contains more discriminative information. They proposed a method to calculate Haar-like features depending on Papageorgiou and colleagues' work [153]:

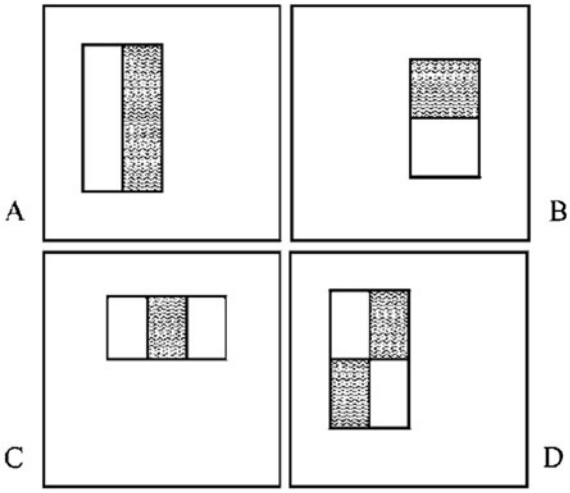

**Fig. 5** Four examples of Haar-like features [32]

$$f = \sum_{i=1}^{k} w^{(i)} \cdot \mu^{(i)} \tag{10}$$

where $f$ is the value of a Haar-like feature, $k$ is the number of rectangles generated from the image, $\mu^{(i)}$ is the mean pixel intensity of $i$th rectangle of the image, and $w^{(i)}$ is the weight of the $i$th rectangle, respectively.

For the weights of the conventional Haar-like features, they are usually fixed and satisfy the following requirements:

$$\sum_{i=1}^{k} w^{(i)} = 0 \tag{11}$$

For example, the weights of three kinds of features used in Viola and Jones' work [32] are sequentially assigned to 1 and −1; 1, −2, and 1; 1, −1, 1, and −1. However, for the extension of Haar-like features in [152], different rectangles are highlighted with different weights. To calculate optimal weights of the enhanced Haar-like features [152], brute-force search (BFS), genetic algorithm (GA), and Fisher's linear discriminant analysis (FLDA) were applied, respectively.

Considering the importance of dynamic characteristics in video-based face analysis, Yang *et al.* [154] designed a dynamic Haar-like feature representation method to extract a set of Haar-like features in a temporal window which contains a sequence of images to reflect different facial expressions. Then, the dynamic feature unit was coded into a binary pattern. By means of binary pattern, the feature vector is changed into a scalar feature, and the encoded feature is robust to noise due to the statistical distribution.

*H. Discussion*

In some HRI scenarios, humans use simple toys to interact with robots. These toys usually have special shapes such as circles or rectangles, and bright colors such as red, yellow, or green. To detect these objects, color and shape features can be used. If these objects are moving, motion features can be adopted. Moreover, attribute-based features can be employed to human-related tasks because human faces have unique colors and textures, and human bodies have special shapes compared with other objects. Generally speaking, for the objects that demonstrate unique characteristics in their color, shape, texture, or motion, attribute-based feature is a good choice to be utilized to represent the objects. In addition, attribute-based features have the advantages like low-CPU use, easy implementation, and simplicity of debugging for developers, hence, they have been widely used in many perception tasks such as object detection and tracking, face detection and tracking, and human detection and tracking that are useful to achieve natural and effective HRI. As discussed above, color features have been successfully applied in object detection and face detection of Kismet and MEXI, and intensity and motion features have been applied in iCub's attention system. While texture and shape features have not been used in the above surveyed robots, they have also been successfully applied in face recognition, object and human detection, and facial expression recognition, and we believe that they can be also employed in social robots.

For algorithm-based features, we briefly introduced DCT-based, Gabor wavelet-based, and Haar-like-based features. Different from attribute-based features, these features are not directly related to the attributes of original signals, but demonstrate better discriminative abilities in some real applications. In real applications, there are many challenging conditions such as large illumination variations and complex backgrounds where the performance of attribute-based features are usually worse. However, under such complex environments, algorithms-based features are more robust to noises, and they have been successfully used in many real world face-related applications such as face detection and recognition, and facial expression recognition.

Among the introduced algorithm-based features, Gabor wavelet-based is the most robust feature because it extracts features from different scales and orientations. Generally, the larger number of scales and orientations, the more information is extracted. Correspondingly, the computational cost is larger. Due to its satisfactory robustness, Gabor wavelet feature still can be used in HRI if its expensive computation problem is solved. Compared with Gabor wavelet features, Haar-like features are more efficient because they do not require complex operations. Hence, they are more popular for real time applications. The famous success is the Haar-like features based real-time face detection. For real applications, how to choose appropriate features in HRI, we should make a balance between the accuracy and the efficiency. Based on this criterion, Haar-like and LBP features are more popular.

As we discussed above, an important objective of HRI is to achieve a natural communication between humans and robots, not only in lab environments, but also in real scenarios. To achieve this goal, more robust features should be extracted for the following semantic understanding tasks.

*2 3D-based*

*(1) Kinect-Based*

Compared with conventional 2D visual signals, Kinect can capture 3D visual signal which contains RGB and depth information simultaneously, also called RGB-D signals. Generally speaking, RGB images are rich in color and texture, and have higher resolutions. However, they are easily affected under non-ideal illuminations. Depth images are robust to illumination changes, but sensitive to low-signal strength returns. Due to the complementary information contained in RGB and depth images, the combination of them can largely increase object detection and recognition, manipulation, navigation, and interaction capabilities in HRI.

Depth images are created by one infrared (IR) projector and one IR camera of Kinect. The value of each pixel in depth image refers to the distance between the object and Kinect. Since such distance can be known in advance and the relations of different objects can be obtained, one simple application using depth image is to provide useful and complementary information to RGB images. Lai *et al.* [155] performed object segmentation by using visual cues, depth cues, and rough knowledge of the configuration between the turntable and camera, simultaneously. Since the distance between the turntable and camera can be represented in depth image, they first removed the background, and the rough position of turntable can be estimated. Due to noise in the depth image, there are usually some small, dark, transparent, and reflective objects on the segmented images. To address this, they exploited a background subtraction method to generate another segmentation. Lastly, both depth-based and vision-based segmentations are combined to better detect objects.

Depth images can be also considered as gray-scale images and the depth values are equivalent to intensity information in conventional 2D images. Hence, the widely used 2D feature extraction methods can be also used for depth images. For example, Benavidez *et al.* [157] applied a 2D gradient and a 2D log filter in depth images to identify the traversable area for their mobile robot. Spinello *et al.* [198] proposed Histograms of Oriented Depths (HOD) to detect people from depth data. Their method was based on the Histogram of Oriented Gradients (HOG) method and the peculiar depth characteristics of the RGB-D sensor.

In addition to 2D feature extraction methods, each depth image can be converted to 3D cloud points by mapping each pixel into the corresponding 3D coordinates. Hence, the existing 3D feature extraction methods can be utilized. Spin Images [158] and Fast Point Feature Histogram [159] are two representative examples. While these feature representation methods have been developed on cloud points and achieved reasonably good performance, they are not suitable for depth images where full viewpoint independence are missed. To address this problem, Bo *et al.* [160] presented five depth kernel descriptors to capture local features to describe the size, shape and edges (depth discontinuities) of objects in depth images for object recognition. To capture the size of an object, Gaussian kernel was applied to compute the similarity between the distance attributes of two cloud points, where the distance is calculated between each point and the reference point. To capture the shape of an object, kernel PCA features and spin kernel descriptors were exploited. For the edges of an object, 2D feature extraction methods such as Gradient and local binary pattern kernel descriptors were used, respectively.

*(2) Stereo-Based*

Stereo vision is another method which can obtain 3D information of objects. Different from depth images where 3D information is directly provided by the depth cameras such as the Microsoft Kinect camera, 3D information from stereo vision can be obtained using multiple images of a single scene from different views. Motivated by the fact that human beings can estimate the distance from objects using two eyes, the simplest stereo vision camera system consists of two separate cameras, and they are commercially available.

Since the technical parameters of stereo camera system are known in advance, the distance information of the object can be computed by the trigonometric measurement method [161] after the object is detected and matched from two images captured by a stereo camera system. Figure 6 shows the object viewed from 2 cameras [162].

In Fig. 6, $d$ is half of the distance between two cameras, $b$ and $c$ are the distances between two cameras and the object, $l$ is the distance between the center of two cameras and the object, $B$ and $C$ are the angles of two cameras, and $A$ is the angle of cameras' cross point. $d$, $B$, $C$, and $A$ are known, and $b$, $c$, and $l$ are unknown. $l$ is the final objective to be achieved. To calculate $l$, $b$ or $c$ should be firstly obtained. The following equations are employed:

$$a^2 = b^2 + c^2 - 2bc \cdot \cos A \tag{12}$$

$$\frac{b}{\sin B} = \frac{c}{\sin C} = \frac{2d}{\sin A} \tag{13}$$

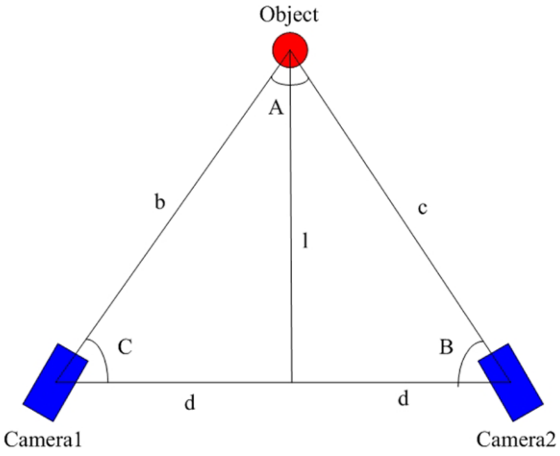

**Fig. 6** Trigonometry

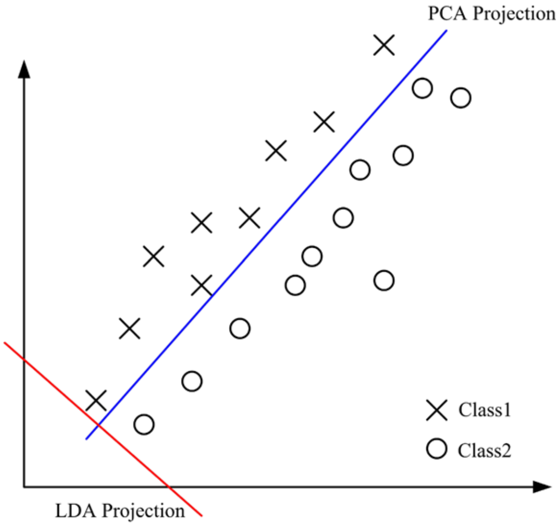

**Fig. 7** Projection directions of PCA and LDA on a toy dataset

Having obtained $b$ and $c$, we can compute $l$ as follows:

$$l^2 = c^2 + d^2 - 2cd \cdot \cos B \tag{14}$$

$$l^2 = b^2 + d^2 - 2bd \cdot \cos C \tag{15}$$

Compared with the depth-based 3D information acquisition, stereo vision has a widespread use in robot-related applications due to the mature techniques and has achieved good performance on acquiring the distance information of objects. For example, Li *et al.* [163] employed stereo vision to estimate the user's attention direction from the head pose and eye gaze direction for their mobile robot. Thompson *et al.* [164] presented stereo vision localization estimation methods to humanoid robots.

*(3) Discussion*

3D visual signals can provide distance information of objects which the robot is manipulating or interacting. With the help of such positional information of objects, the robot will know where the object is and how the surrounding environment looks like. This will better improve the robot's operation performance and make the robot more human-like. As described above, both Kinect and stereo vision systems can obtain 3D information of objects.

There are some advantages and disadvantages for the two 3D vision systems. For the Kinect vision system [155–157, 160, 198], the camera utilizes an infrared pattern that is projected on the scene and can directly acquire the depth image to provide the distance information of objects. Depth images also can be converted to 3D cloud points by mapping each pixel into the corresponding 3D coordinates. Meanwhile, the Kinect sensor provides RGB information. Hence, researchers can conduct their developed approaches on both the RGB and depth images. While the Kinect camera demonstrates good performance in many indoor environment applications, the performance is generally poor when applied in outdoor environments. Moreover, the Kinect camera can only work over a range of several meters, which is another limitation for real-world applications.

For the stereo vision, many commercially available systems have been developed due to mature techniques. These systems are of different sizes and can be used for both indoor and outdoor environments [161–164]. Moreover, it can produce higher resolution than Kinect. However, since 3D information of a stereo vision system is usually calculated by using at least two images of the object captured from different views, it heavily depends on the results of feature detection and correspondence. Even if many commercial and free stereo libraries are available, the performance of current stereo vision systems will also be affected especially for real-time and real-world applications due to several challenging problems of 2D image processing like illumination changes. Therefore, with respect to 3D information acquisition, the users should choose proper systems according to their own requirements.

### *3.1.2 Audio-Based Methods*

Audio has also been widely used in many social robots to implement semantic understanding tasks, such as speaker localization, speech recognition, sound event classification, emotion recognition, and rhythms recognition. In this paper, we will not describe audio features for speech recognition because a number of speech recognition software solutions are commercially available, such as IBM's ViaVoice, Dragon Naturally Speaking, Voice Xpress, FreeSpeech 2000 from Philips [165] and CMU Sphinx. They can be directly used in social robots. For instance, Simmons *et al.* [27] selected IBM's ViaVoice to convert spoken utterances to text strings, and Esau *et al.* [51] chose ViaVoice in MEXI. Similar to visual-based methods, different semantic understanding tasks use different features.

While audio features for speech recognition will not described in this paper, we will briefly introduce phoneme detection from a raw audio signal since it is important for several audio related applications such as emotion recognition that use phoneme detection to parameterize the audio signal and account for prosodic differences. Phoneme detection normally includes three stages: feature extraction, segmentation and labeling, and word-level recognition [166]. First, to describe the broad acoustic properties of the different phonetic units, feature detection with a spectral analysis of the speech is applied to convert the spectral measurements to a set of features. Next, a segmentation and labeling phase is used to segment the speech signal into stable acoustic regions. Following that, attaching one or more phonetic labels to each segmented region is implemented, that could result in a phoneme lattice characterization of the speech. Last, it should match a valid word with the phonetic label sequences [167].

To recognize speech-based emotions, acoustic features such as prosodic, voice quality, and spectral features have been widely used because it is easily to extract and process these features, even in distinct cultural backgrounds. As Clavel *et al.* [168, 169] stated, prosodic features mainly include fundamental frequency, intensity contours and durations; voice features normally contain jitter, shimmer, unvoiced rate, and harmonic to noise ratio; and spectral features usually consist of Mel Frequency Cepstral Coefficients (MFCCs), Bark band energy, and spectral centroid. In addition, other features were also presented for acoustic recognition, such as phoneme dominant feature [170], and low-level time and frequency derivatives [37]. Please refer to [171, 172] for details.

In MEXI [51], an emotion recognition system—PROSBER was developed, in which six speech parameters including the fundamental frequency, energy, jitter, shimmer, power spectrum, and speech/pause time were exploited. In BARTHOC Jr [173], to recognize emotion states, 1316 features on pitch, energy, MFCCs, frequency spectrum, duration, and pauses were computed, and 20 features related to pitch, MFCCs and energy were selected to further improve the performance and speed. In Kismet [22], 12 kinds of audio-based features related to pitch and energy including pitch mean, pitch variance, maximum pitch, minimum pitch, pitch range, delta pitch mean, absolute delta pitch mean, energy mean, energy variance, energy range, maximum energy, and minimum energy were applied to identify approval, attentional bid, prohibition, soothing, and neutral emotions.

Among the above mentioned audio features, pitch, energy, and MFCCs are three representative ones. Pitch refers to the fundamental frequency which defines the lowest frequency when the speech signal repeats. Pitch contour, range, mean, median, inflection range, and rate of change are related to distinct emotions [174]. To extract pitch features, several methods were developed, such as the simplified inverse filter tracking (SIFT) [175], the pitch constraints and dynamic programming search [176], and the statistical method based on cepstrum [177].

Energy is defined as the volume or intensity of speech which aims to measure the variations of speech signals' amplitude. The energy is usually computed as the log of the signal energy [178], described as below:

$$E = \log \sum_{n=1}^{N} S_n^2 \tag{16}$$

where $S_n$ is audio sample, $n = 1, \dots, N$, and $N$ is the number of audio samples.

MFCCs are coefficients which collectively make up an mel frequency cepstrum (MFC) to represent a short-term power spectrum of a sound. The MFCCs are calculated as [178]:

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{K} (\log S_k) \cos\bigl[n(k-0.5)\pi/k\bigr] \tag{17}$$

where $S_k$ is the output of the filter bank which includes 12 triangular filters, $n = 1, 2, \dots, N$, and $N$ is the number of audio samples.

Based on the basic acoustic features described above, some researchers have exploited advanced features to reflect the statistical performance and attributes of the raw audio signals. For instance, Vogt and André derived features including the minima and maxima, temporal distance, the first and second derivatives, *etc.* based on pitch, energy and MFCC, and their statistical parameters were further computed as final feature vectors, such as mean, variance, median, *etc*. Some other features like positions of the overall pitch maximum, the number of pitch, and energy minima and maxima per segment were also adopted for feature representation. Please refer to [59] for details.

In addition, Kim *et al.* [179] proposed a novel speaker-independent feature, the ratio of a spectral flatness measure (SFM) to a spectral center (RSS), to recognize emotion. The RSS is calculated as shown in Eq. (18). SFM is the ratio of the geometric mean to the arithmetic mean of the power spectrum, and the spectral center is the average frequency weighted by acoustic power, computed as shown in Eqs. (19)–(20), where $X_j(f)$ is the magnitude of the short-term Fourier transform with the $f$th frequency for the $j$th speech window with length of $N$. Both of them are related to the extent to which the utterance is noise-like or tone-like. The presented features have been used in the emotion recognition system developed in a Korea robot to help the elderly and lonely people.

$$RSS_j = \frac{1000 \times SFM_j}{spectral \quad center_j} \tag{18}$$

$$SFM_j = \frac{(\prod_{f=1}^{N} X_j(f))^{1/N}}{1/N \sum_{f=1}^{N} X_j(f)} \tag{19}$$

$$Spectral \quad Center_j = \frac{\sum_{f=1}^{N} f_j \cdot X_j(f)}{\sum_{f=1}^{N} X_j(f)} \tag{20}$$

There is another kind of spatio-temporal box filters (STBF) feature to characterize periodic sampling features and temporal integration features based on mean, min, max, standard deviation, and quadrature pair operations. The extracted features represent critical properties of audio signals. In Ruvolo and colleagues' work [37], to recognize emotions, six types of box filters (also called Haar-like features) were used in a 2D Sonogram, which was converted from the raw audio signals. The reason to use Sone units is that they are proportional to human's loudness.

Besides emotion recognition, sound localization is another representative semantic understanding task based on audio signals. Correspondingly, the specific features are extracted to accomplish this task. For example, three features called interaural time difference (ITD), interaural level difference (ILD), and notches in the frequency response of each ear were proposed in sound localization, respectively [18]. ITD is the ratio of the number of samples to the sampling frequency. The cross-correlation function was executed on the signals from left to right microphones, and the sampling frequency is pre-defined. For ILD, it is computed as a function of the average power of the audio signals. The used signals were temporal-based. To obtain notches, frequency notches were calculated and more steps were taken. First, power spectra density for each ear was computed through the Welch spectra method [180]. Then, interaural spectral difference (ISD) was calculated. Based on the obtained results, a 12-degree polynomial curve was fitted on them. Lastly, minima of the fitted curve was found as the notches feature. Shibata and colleagues have applied ITD and IID features in their robots for sound localization [181].

Motivated by the fact that speech is an important human–human communication channel, audio signals also play significant roles on HRI of social robots. By analyzing the collected audio signals, robots can acquire more information related to their interaction subjects, such as their positions, commands, and emotional states. These information are useful and helpful to build a natural communication between humans and robots. To successfully and accurately obtain these information, extracting effective audio features is a key factor.

Through reviewing the commonly used features in emotion recognition and sound localization, it can be seen that how to develop audio features mainly depends on semantic understanding tasks. For example, for speech-based emotion recognition, three groups of features including prosodic,

**Table 4** Performance comparison of visual- and audio-based low-level features in perception. Generally, there is not a clear criterion to define these qualitative results. Different from some quantitative results such as precise accuracy and computational cost, we just present a general concept of these qualitative performance of different methods in this table. Specifically, these results are obtained by comparing these methods together to obtain the relative qualitative results

| Method | Feature | Accuracy | Computational complexity | Robustness |
|---|---|---|---|---|
| Visual-based | color [94–96] | medium | low | low |
| | intensity [99–101] | high | medium | medium |
| | texture [104–106, 110] | high | high | high |
| | shape [116–119] | low | medium | low |
| | motion [120–123] | medium | high | medium |
| | DCT [146, 147] | high | high | medium |
| | Gabor wavelet [148–151] | high | high | high |
| | Haar-like [152–154] | high | medium | medium |
| Audio-based | pitch [174–177] | medium | medium | medium |
| | energy [178] | low | low | high |
| | MFCCs [178] | high | medium | medium |

voice quality, and spectral features have been widely used. The number of features can be determined by the recognized emotional states and applied environments. This is because each emotional state can be described by different features. Pitch, energy, and MFCCs are three popular feature representation methods in this application. For sound localization, feature selection is similar to that used in speech-based emotion recognition. It depends on the category and environments of localized sounds. For example, due to different characteristics of different sounds, the features to localize the sound of voices and the sound of hand clap are different. Generally speaking, the application scenario of social robots is an important factor to choose and develop audio-based features.

In summary, we present a brief performance comparison of visual- and audio-based features in perception, and the advantages and disadvantages such as accuracy, computational complexity and robustness are compared in Table 4. We briefly analyze why these features are compared as shown in this table.

- In terms of the accuracy, most visual-based features can obtain an acceptable recognition performance. That is because visual features generally contain many discriminative information and most feature extraction methods can exploit these information in perception. Among these features, the performance of shape, color and motion features is generally poorer than others. The reason is that these features mainly focus on global information of visual data and may ignore some local discriminative information. For audio-based features, MFCCs is the best feature representation method and has been widely used in many audio event analysis tasks. That is because it can reflect more detailed information of audio data than the other two features.
- The computational cost of most visual feature extraction methods is generally higher than those of audio-based methods. The reason is feature dimensions of visual data are generally larger than those of audio data.
- In terms of the robustness, texture and Gabor wavelet-based methods are top two visual features because local information of visual data can be effectively extracted by these two methods and hence they are robust to noise.

For the features listed Table 4, once the required parameters are selected, they can be extracted in fully autonomous status. In other words, when social robots employ these features in the perception systems, human interference is not required. However, when the social robots are applied in other environments, the parameters need to be re-selected, and the perception system will be partially autonomous. The advantages and disadvantages listed in Table 4 can be used as a reference for feature determination.

Now, we briefly discuss the difference of HRI and human–compute interaction (HCI). HRI is a growing field of research and application. The field includes many challenging problems and has the potential to produce solutions with positive social impact. Its interdisciplinary nature requires that researchers in the field understand their research within a broader context. Specifically, HRI is closely related to HCI. However, there are at least three important differences between HCI and HRI, described as below [223–225]:

- The main difference between HCI and HRI is that HRI deals with embodied and situated agents. What this means is that perception is grounded in the real world. The consequences for both the sensation and interpretation of events is that for embodied and situated agents they need to be tied to action. For example, perceiving objects in the world is not just about correctly identifying them in an image, but also translating that into knowledge of where that object is relative to an actuator, such as a gripper. Lighting conditions change as the camera moves, and it is not just important to know the u,v coordinates of where an object is in an image, but also the full coordinate of where it is in the world. Finally, when dealing with human interaction, it is also important to balance recognition accuracy with latency in order to effectively sustain an interaction.

- HRI focuses on the combination of the user and the robot and HCI focuses on the user using the computer, which indicates that the main objectives of HRI and HCI are different. The main reason is that several assumptions such as clean environments, controlled conditions, and simple background are made to design the HCI methods, however, these assumptions may not hold in many real applications where HRI is employed.
- The interactions of HCI and HRI are generally different. Specifically, HRI is physically bidirectional because robots are not passive entities like computers and it could implement the functions related to actions or motion; HRI is asymmetric because robots have not the same cognitive skills of humans; HRI is unique because robots are perceived as living entities. HRI has lower repeatability because there are even not two robots following the same path in HRI. While HCI has been studied for many years, tools and metrics do not directly transfer to HRI and many methods developed in HCI may not always be readily applicable in the design and evaluation of interactive robots for HRI.

While there are several important differences between HCI and HRI, to our best knowledge, many state-of-the-art HCI methods could be applicable to HRI if the gap between the theoretical research in HCI and the practical application in HRI is bridged. As the field of HRI has grown, it has seen many contributions from researchers in HCI and it has been nurtured by HCI organizations. HRI research is attractive to many members of the HCI community because of the unique challenges posed by the field. HRI benefits from contributions from HCI researchers, both in methodologies, design principles, and computing metaphors.

### *3.1.3 Tactile-Based Methods*

Besides audio- and visual-based features, tactile-based feature is also important in human–robot interaction. It is well-known that when humans communicate with a robot, it is inevitable to interact with the robot physically, which is called tactile-based interaction. Similar to visual- and audio-based signals, tactile signals contain rich information and can be used to environment understanding. For example, a robot can know whether there are persons touching on it and what kind of touches by analyzing the collected tactile signals. While human touch's meanings are still yet to be explored, it can reflect the internal state of humans to a certain extent. For instance, if humans lightly pat the robot, it may show that they are be in a relaxing state; if they heavily scratch the robot, it may mean that they are in a fretful state. Moreover, a robot can acquire information of the touched objects by exploring them with various kinds of tactile sensors [182]. Generally, there are two popular methods including singular tactile sensor-based and "sensitive cover"-based. Singular tactile sensor-based is usually embedded in robots' arms and grippers, and "sensitive cover"-based is to detect a full body's or region's sense of touch.

Dao *et al.* [182] designed a 4-DOF (degree of freedom) soft-contact tactile (SCT) sensor which can detect three components of force and one component of moment. It has been employed in robots' fingers to sense the gripping forces when gripping objects. However, when social robots interact with users, they sense human's touch not only through fingers, but also through the whole body, especially for robots in the form of toys. Under such situation, "sensitive cover"-based interaction is applied. Tsetserukou *et al.* [183] used "sensitive cover" in a robot arm to automatically generate compliant motion according to the measured external force vector. Here, tactile sensors and torque sensors were distributed in the "sensitive cover". Iwata *et al.* [184] employed "sensitive cover" to another robot arm that was able to avoid physical interference from human. The "sensitive cover" was designed by using force-torque sensors and touch sensors. Gorostiza *et al.* [83] designed a personal robot named Maggie that can interact with humans in a peer-to-peer relationship. The tactile system of Maggie consists of several invisible tactile sensors installed in the robot's casing. Maggie can detect tactile events from human. Stiehl *et al.* [185] presented a "sensitive skin" containing three kinds of sensors under a soft silicone skin to detect external force and temperature. It has been applied in Huggable [50, 186], a new type of therapeutic robotic. Shibata [187] developed a ubiquitous surface tactile sensor that can collect both position and pressure information. Moreover, the designed sensor system has advantages of flexibility and usability to a curbed surface. It has been employed in the therapeutic robotic, Paro.

As described above, "sensitive cover" is a popular way to detect forces from the outside environments. The key component of "sensitive cover" is the tactile sensor matrix. Since each sensor in the matrix can generate a pressure value, a number of sensors can generate a tactile image representing the pressure distribution. Different from a visible image, a tactile image is usually of low-resolution and contains some mechanical cross-talk noise [188]. Even so, it can also be used for semantic understanding tasks.

Besides 1D tactile signals, 2D tactile images have also been used to extract tactile-based features. For the raw tactile signals, Iwata *et al.* [189] extracted 11 parameters as the input of a neural network classifier to recognize touching states, such as hit, grasp, scrape, and scratch. The extracted parameters include maximum force, time to reach the maximum force, contact duration, number of contacts, average deviation of force, average force, maximum pressure, maximum contact area, maximum two-point distance, moved distance of contact area, and summation of contact area, which were calculated by the conventional statistical methods of 1D signals. Stiehl and colleagues [186] extracted features such as normalized sensor value sum, normalized average sensor value sum, the change in sensor value sum, and the centroid location and direction of motion, in the Huggable robot. In addition, Göger and colleagues [190] applied PCA to classify slip and itons between the robot gripper and the object which has to be manipulated. Before executing PCA, tactile signals were transformed from the time domain to the frequency domain by using a short-time Fourier transform defined as $STFTx[n] \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n]\omega[n-m]e^{-i\omega n}$, where $x[n]$ is the raw signals, $\omega[n]$ is the window used. Then, an image with 1024 pixels was formed based on the computational results of eight FFT transformations. Finally, PCA was conducted on a dataset containing $n$ images.

For tactile images, we can process them similar to the conventional visible images. For instance, each pixel in a tactile image called taxel and can be calculated similar to the intensity in visible images. Differently, taxel values are calculated in different ways due to the exploited tactile sensory system. For example, Payeur *et al.* [184] took 96 “taxels” (pixels with pressure force value instead) generated from an $8 \times 12$ tactile image as the input of a two-layer feed-forward neural network to recognize the shape of an object. Here, tactile values were represented by geometric displacements. Similarly, Carotenuto *et al.* [191] directly employed tactile images to determine objects’ category where taxel values were the sum of the normal stresses. In addition, feature extraction methods used in visible images can be extended to tactile images. An example is Göger and colleagues [190] who applied PCA to tactile images to identify contact patterns.

For tactile-based feature extraction methods, most conventional signal processing can be employed, such as the methods to extract statistical features and short-time Fourier transform, PCA and Active Hough Transform (AHT). In D. Berger’s work [188], after removing the effects of cross-talk noise from the tactile images by using a adaptive thresholder, an edge detector was used to detect edges and AHT was deployed to determine the line parameters.

In terms of the popularity of the above feature extraction methods, tactile-based methods are similar to visual- and audio-based methods. Moreover, tactile-based methods can be further used when the robots only have physical contacts with the outside environments. As we mentioned above, “sensitive cover”-based methods are more popular. That is because they can not only detect point contacts, but also detect surface contacts which usually happen in HRI. However, the price of “sensitive cover” is comparatively high.

### *3.1.4 Range Sensors-Based Methods*

Besides human-like methods, social robots can use other perceptual systems to understand the outside environments, e.g., laser range finders. Laser range finders are commonly used to detect paths or obstacles in mobile robots for navigation. Since only human–robot interaction is surveyed in this paper, we concentrate its applications on tracking people and predicting people’s behaviors. Considering the executed tasks, there are two ways to place laser range finders, one is to mount them in the mobile base of robots, and the other is to place them in the environments where robots act.

Kanda *et al.* [35] selected the second mounting method to place six SICK LMS-200 laser range finders in a shopping arcade where they acquire humans’ trajectories to predict their behaviors, such as people’s walking styles and directions. These information of human behaviors has been exploited by Robovie. To obtain human trajectories, Particle Filtering was used to track torso-level scan data by a background substraction model. The background model was obtained by filtering out noise and objects from hundreds of frames. Then, points satisfying some pre-defined rules were identified as persons. By using customized calculation rules, 32 features were extracted to represent the end point of the normalized trajectory, the size of the rectangle that covers the normalized trajectory, the angles of the trajectory, and the velocity for local behavior prediction. The proposed method was first applied in an experiment room to track human body’s motion, and the centers of ellipses with axis between 15 cm and 55 cm were identified as human candidates [192].

For the first mounting way of laser range finders, the collected data are used to track people and provide additional information. For human tracking, a challenging problem is how to localize people’s initial positions, and there are several methods proposed in [35] and [192]. The similar idea was also used in GRACE [193] which is to utilize the size information of the segmented regions to detect humans. The segmentation with the width between 5 cm and 60 cm was classified into potential person category or human leg category. In addition, Jensen *et al.* [9] presented a method to detect people by using the motion information. First, a static map containing non-moving objects was generated according to the laser readings. Then, new information was obtained by comparing new readings and the static map. Thirdly, a chi-square test was applied to the new information to determine whether it was the static map or a moving object. Lastly, the static map was updated for a new validation. The information labeled as dynamic objects was human candidate.

Jung and Sukhatme exploited a laser range finder to provide depth information for estimating the partial 3D position [194]. Given the optical properties of a camera, distance

**Table 5** Summary and comparison of different categories of low-level feature extraction methods in social robots. Specifically, these results are obtained by comparing these methods together to obtain the relative qualitative results

| Method | Accuracy | Computational complexity | Robustness | Cost | Convenience |
|---|---|---|---|---|---|
| Visual-based [94, 99, 105, 144, 152] | medium | high | medium | low | high |
| Audio-based [174–178] | medium | medium | low | low | high |
| Tactile-based [189–191] | high | high | high | medium | medium |
| Range sensors-based [192–195] | low | medium | medium | high | high |

information from the laser range finder can be projected onto the image coordinates through a transformation equation as:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{w}{2} \times (1 - \frac{\tan(\alpha)}{\tan(f_h)}) \\ \frac{h}{2} \times (1 + (d - \frac{d}{r} \times (r - l) \times \frac{1}{l \times \tan(f_v)})) \end{bmatrix} \tag{21}$$

where $(x, y)$ represents the coordinate of each pixel in the image, $\alpha$ and $r$ are the heading and the range of a scan, respectively, $f$ is the focal length of the camera, $f_h$ and $f_v$ are the horizontal and vertical angles of view of the camera, $d$ is the height from the laser range finder to the camera, and $w$ and $h$ are the width and height of the image. Additionally, Glas and colleagues deployed laser scan data to estimate human positions by providing occupancy and edge information. Occupancy information represents the status of a coverage grid: occupied or empty represented by integrating values of all sensors, and edge information indicates edges of an identified object represented by the edge-fit estimation using an angular array. More details can be found in [192, 195].

Besides the above introduced 2D range sensors-based methods, 3D laser range finder has been developed to collect 3D laser reading information to explore the whole 3D scene around the robot. The basic idea to design a 3D range finder is to mount a 2D range finder onto the rotating support such that the scan measurement can perform horizontal scan and vertical scan by two scanning mechanisms [196]. Similar to the stereo vision system, 3D range finder could be commercially available according to the specific requirements.

To extract features from 3D data, there are usually two kinds of methods: the first is to use appropriate feature extraction methods on 3D data directly, and the other is to transform 3D data into 2D form and then utilize existing 2D methods to extract features. Scholer *et al.* [197] employed the observation model of a particle filter to 3D laser range data to track a person with occlusions. Spinello *et al.* [198] utilized 3D range data to detect people. Different from Scholer's work, they treated the 3D cloud points as a collection of 2D laser points at different slices or layers. For 2D laser data at each layer, the authors extracted 17 existing 2D features, such as width, circularity, PCA ratio, quadratic spline fitting, bounding box area and so on. Navarro-Serment *et al.* [199] exploited 3D LADAR data to detect and track people. In their work, they applied both 2D and 3D feature extraction methods from 3D cloud points of the object. First, they projected 3D point cloud into 2D planes by using a virtual scan line. Then four motion-related variables including the object's size, the distance it traveled, and the variation in the object's size and velocity were computed. By utilizing the position and size information of the object, 3D point sets referring to potential humans can be isolated. Then, for each potential object, geometric features including 2D-based and 3D-based were extracted. Among them, the covariance matrix, and the normalized moment of inertia tensor were calculated by using 3D cloud points.

While laser range finders have been widely applied in mobile robots, they require the robot to have the ability to scan the environments when it moves at a reasonable pace. Under such scenario, it is more difficult to collect high quality 3D cloud points by using 3D laser range finder compared with 2D lasers, especially in outdoor environments [200]. Therefore, if the corruption of the 3D data cannot be repaired in the data collection, the performance of these used feature extraction methods are generally heavily affected, and more robust feature extraction methods should be used.

To better compare the different performance of the above feature extraction methods, we list their advantages and disadvantages in Table 5, where accuracy, computational complexity, robustness, practical cost, and convenience are tabulated. We can see that different feature extraction methods demonstrate different performance in terms of different factors. We present the following reasons to discuss why they are summarized as shown in the table.

- In terms of the accuracy, tactile-based method is the best and range sensors-based method is the worst. That is because tactile-based methods are based on physical touches and robust to external disturbances, and lasers are sensitive to these disturbances. For visual- and audio-based methods, their accuracies are medium in general. For example, the accuracies of Gabor wavelet [148–151], texture [104–106, 110], and Haar-like visual-based features are generally higher and those of intensity and shape features are lower. Similarly, the accuracy of energy audio-based feature is high and those of pitch and MFCCs are low.

- The computational complexity of visual- and tactile-based methods is generally higher than the other two methods. The reason is the data dimensions of signals of these two features are generally larger than those of the others, and hence a higher computational complexity is usually required.
- In terms of the cost, audio and visual sensors are much cheaper than lasers and tactile-based sensors and they have already been widely used in our daily lives. Compared with tactile-based sensors, range sensors-based sensors are most expensive and hence are only suitable to specific application tasks.
- For the convenience of use, tactile-based method is generally worse than the other three because it requires physical touches in practical applications. Therefore, it is difficult to achieve uncontactness HRI for this method.

### *3.1.5 Multimodal-Based Methods*

The above described feature extraction methods are unimodal-based. While some encouraging performance can be obtained, there are still several shortcomings because unimodal-based features are generally not robust in many real-world applications. To address this problem, multimodal feature extraction methods are usually used in many social robots. We here present three representative exemplars in the following.

The first example is the attention system of robotics, where multimodal signal fusion is applied. For example, Matthias *et al.* [16] designed iCub's attention system via audio-visual synchrony for tutoring a child. To effectively detect the synchrony, mutual information was utilized to measure the correlation between visual and audio signals. For the visual signal, image intensity feature was used, and for the audio signal, audio energy was applied. In iCub, another method related to the attention system was also presented [99]. It fulfilled a multimodal saliency-based bottom-up attention. Intensity, color, directional and motion information were visual saliency features, and center location, uncertainty information, and intensity values were auditory saliency features. To obtain a multimodal saliency map, all visual and audio saliency features were converted to a common egocentric reference and aggregated into a single map by choosing the maximum value from all saliency channels at the same position. The highest saliency values in the combined map were taken as the attentions of the robot. By using the designed attention systems, iCub can detect objects of interest.

Another example is the audiovisual-based emotion recognition. As described above, facial and vocal signals are two useful and widely-used sources to detect and characterize human emotions. However, there are some shortcomings in unimodal-based affective state recognition [201]. For example, facial expression recognition is easily affected by pose and illumination variations, and noise is a significant effect on audio signal analysis. To address these issues, some researchers have pointed out that audiovisual-based (fusion of audio and visual signals) emotion states recognition is a new solution to improve the performance. For example, Fragopanagos *et al.* [202] designed an Emotionally Rich Man-machine Intelligent System (ERMIS) to recognize emotions, where Artificial Neural Network for Attention (ANNA) was used to fuse features from different modalities, such as linguistic, paralinguistic (prosodic) and facial images. Zeng *et al.* [203] used Multistream Fused Hidden Markov Model (MFHMM) to detect and track a user's emotion states by using both audio and visual signals.

The last example is [204]. They proved a hypothesis that HRI through semantic integration of multiple modalities, dialog management, and contexts can show better performance than that obtained from a single modality. A domestic robot to clean a house using a multi-modal interface was applied to test the hypothesis. The multi-modal interface included audio modal like speech, and visual modal such as pointing, field of vision, and head nodding. The compared single modal is the speech. Accuracy and succinctness are two criteria to evaluate the robot's learning process. After a series of experiments, they concluded that multiple modalities can improve HRI over single model and the robot can better implement the specified tasks than using a single modality.

While these exemplars have shown that multimodal-based feature fusion methods usually perform better than unimodal-based methods, there are still several unsolved problems. For example, which fusion level, feature level or decision level, is better? How these methods work if one source of feature is very noisy? How to handle large scale multimodal features and combine multimodal information with different representations still need more investigations in future work.

## 3.2 Dimensionality Reduction

Features extracted from different modalities are usually high-dimensional and there may be some redundancy in these features. Hence, it is desirable to use some statistical learning techniques to further process these features to better reflect their semantic information. Subspace learning is one of such representative methods. Subspace learning aims to find a mapping to project high-dimensional raw signal into a low-dimensional feature space, such that some intrinsic characteristics of the original signals can be revealed and preserved. Representative methods include principle component analysis (PCA), linear discriminant analysis (LDA), and locality preserving projections (LPP), respectively [127–142].

Considering a set of samples denoted as a vector-represented dataset $\{x_i \in R^d, i = 1, 2, \ldots, N\}$ and the corresponding label $\{l_i \in \{1, 2, \ldots, c\}, i = 1, 2, \ldots, N\}$, where $N$ is the number of samples, $d$ is the feature dimension of each sample, $c$ is the number of classes, and $x_i$ possesses a class label $l_i$. The objective of subspace learning is to find a linear mapping $W = [w_1, w_2, \ldots, w_k]$ to project $\{x_i, i = 1, 2, \ldots, N\}$ into a low-dimensional representation $\{y_i \in R^m, i = 1, 2, \ldots, N\}$, i.e., $y_i = W^T x_i$, $m \ll d$. The essential difference of these subspace learning methods lies in the difference in defining and finding the mapping $W$.

PCA seeks to find a set of projection axes such that the global scatter is maximized after the projection of the samples. The global scatter can be characterized by the mean square of the Euclidean distance between any pair of the projected sample points, defined as [143]

$$J_T(w) = \frac{1}{2}\frac{1}{NN}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - y_j)^2 \tag{22}$$

We can simplify $J_T(w)$ to the following form

$$\begin{aligned} J_T(w) &= \frac{1}{2}\frac{1}{NN}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(w^T x_i - w^T x_j\right)\left(w^T x_i - w^T x_j\right)^T \\ &= w^T\left[\frac{1}{2}\frac{1}{NN}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(x_i - x_j)^T\right]w \end{aligned} \tag{23}$$

Let

$$S_T = \frac{1}{2}\frac{1}{NN}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(x_i - x_j)^T \tag{24}$$

and the mean vector $m = \frac{1}{N}\sum_{i=1}^{N} x_i$, then $S_T$ can be calculated as follows:

$$S_T = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)(x_i - m)^T \tag{25}$$

then Eq. (22) can be rewritten as

$$J_T(w) = w^T S_T w \tag{26}$$

The projections $\{w_1, w_2, \ldots, w_k\}$ that maximize $J_T(w)$ comprise an orthogonal set of vectors representing the eigenvectors of $S_T$ associated with the $k$ largest eigenvalues, $k < d$, which is the solution of PCA.

LDA seeks to find a sets of projection axes such that the Fisher criterion (the ratio of the between-class scatter to the within-class scatter) is maximized after the projection. The between-class scatter $S_B$ and the within-class scatter $S_W$ are defined as [144]

$$S_B = \sum_{i=1}^{c} N_i(m_i - m)(m_i - m)^T \tag{27}$$

$$S_W = \sum_{i=1}^{c}\sum_{j=1}^{N_i}(x_{ij} - m_i)(x_{ij} - m_i)^T \tag{28}$$

where $x_{ij}$ denotes the $j$th training sample of the $i$th class, $m_i$ is the mean of the training sample of the $i$th class and $m$ is the mean of all the training samples. The objective function of LDA is defined as

$$\max_{w} \frac{w^T S_B w}{w^T S_W w} \tag{29}$$

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

$$S_B w = \lambda S_W w \tag{30}$$

Let $\{w_1, w_2, \ldots, w_k\}$ be the eigenvectors corresponding to the $k$ largest eigenvalues $\{\lambda_i | i = 1, 2, \ldots, k\}$ decreasingly ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, then $W = [w_1, w_2, \ldots, w_k]$ is the learned mapping of LDA. Since the rank of $S_B$ is bounded by $c - 1$, $k$ is at most equal to $c - 1$. Figure 7 shows the projections of PCA and LDA on a toy dataset.

LPP [145] is one recently proposed manifold learning method, and the aim of LPP is to preserve the intrinsic geometry structure of original data and make the samples lying in a neighborhood to maintain the locality relationship after projection. Specifically, LPP first constructs one affinity graph to characterize the neighborhood relationship of the training set and then seeks one low-dimensional embedding to preserve the intrinsic geometry and local structure. The objective function of LPP is formulated as follows:

$$\min \sum_{ij}(y_i - y_j)^2 S_{ij} \tag{31}$$

where $y_i$ and $y_j$ are the low-dimensional representation of $x_i$ and $x_j$. The affinity matrix $S$ can be defined as:

$$S_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2/t), \\ \quad \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ 0 \quad \text{otherwise} \end{cases} \tag{32}$$

where $t$ and $k$ are two dependent pre-specified parameters defining the local neighborhood, and $N_k(x)$ denotes the $k$ nearest neighbors of $x$. Following some simple algebra deduction steps [145], one can obtain

$$\frac{1}{2}\sum_{ij}(y_i - y_j)^2 S_{ij} = wXLX^T w \tag{33}$$

where $L = D - S$ is the Laplacian matrix, $D$ is a diagonal matrix and its entries are column sums of $S$, i.e., $D_{ii} = \sum_{ij} S_{ji}$. Matrix $D$ provides a natural measure on the data points, and the bigger the value $D_{ii}$ is, the more "important" $y_i$ is. Usually, one can impose a constraint:

$$y^T D y = 1 \quad \Rightarrow \quad w X D X^T w = 1. \tag{34}$$

Then, this minimization problem can be converted into solving the following constrained optimization problem:

$$\begin{aligned} & w_{opt} = \arg\min_{w} w^T X L X^T w \\ & s.t. \quad w^T X D X^T w = 1. \end{aligned} \tag{35}$$

Finally, the bases of LPP are the eigenvectors of the following generalized eigenvalue problem:

$$X L X^T w = \lambda X D X^T w \tag{36}$$

Let $w_1, w_2, \ldots, w_k$ be the solutions of Eq. (31), ordered according to their eigenvalues, $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$, then $W = [w_1, w_2, \ldots, w_k]$ is the resulted mapping of LPP.

## 3.3 Semantic Understanding

Having extracted the features, semantic understanding tasks should be implemented for social robots. Generally speaking, different tasks involve different semantic understanding methods. For example, object recognition, emotion recognition and human identification need classification methods, and face and human tracking requires tracking techniques. Hence, we discuss existing approaches which have been used in several typical applications of HRI such as object recognition, object tracking, object segmentation, and speaker localization, and show their characteristics and limitations in this subsection.

### *3.3.1 Object Recognition*

Object recognition is one of the most representative applications in HRI, such as face recognition, human identification, and emotion recognition. In this subsection, we unify them into a general object recognition framework,[7] and present state-of-the-art object recognition methods in the following.

*A. Template Matching* Template matching is one of the simplest recognition methods, and the basic idea is to recognize the object to a gallery sample which has the highest similarity. For example, Scassellati [13] applied ratio template matching in Cog to detect human faces, in which a 14×16 gray-scale image patch was chosen as a template, and the ratio of average gray-scale value was utilized in the template region. If the ratio exceeds a pre-defined value, the corresponding candidate region was detected as a face region. Moreover, to speed up the method, an early-abort scheme and a motion-based pre-filter were presented. Similarly, the ratio template matching approach can be also used for eye detection.

*B. Clustering* Clustering is another machine learning method to gather similar objects. Hasanuzzaman *et al.* [34] used a multi-cluster approach to classify face and hand poses based on PCA-based features. In their work, Euclidean distance was applied to measure the similarities of different images. Nickel and Stiefelhagen [46] applied a $k$-means clustering method to identify face and hand regions. Differently, they took skin color and image disparity as features. Before clustering, skin-color probability histogram features were calculated. Then, dilation and erosion operations were applied on the skin-color map. Simmons and colleagues [27] applied a clustering method on range readings from SICK scanning laser range finders to identify people and walls. The shape of the cluster was a criterion for judgement. A small cluster and a pair of small clusters were candidates of people, and big clusters having line shapes were candidates of walls.

*C. Nearest Neighbor* Many object recognition methods use the nearest-neighbor classifier, such as face recognition in ARMAR III [147]. It first computed the distances between the test image and the training images, and applied a Min-Max normalization method and a sum rule to normalize and fuse the scores. Then, the face was recognized according to the highest score obtained and a predefined threshold value. It has been also used in Göger and colleagues's work [190] to classify slip status including signal with slip, signal without slip, and noise and contact patterns like small point contact, large point contact, edge contact *etc*.

*D. Neural Network* Neural Network is a well-known classification method, which consists of several neurons interconnected with each other to form input, hidden and output layers to stimulate human's Neural Network. Breazeal and colleagues [20] selected a pixel based multi-layer perceptron to recognize areas of eyes and nose bridges. For the perceptron's inputs, they are subimages including eye candidates extracted from potential faces. Voit *et al.* [45] applied Neural Network to estimate head poses. The neural network used includes three layers, and normalized intensity image and the Sobel magnitude were fed into the input layer. For the hidden layer, 100 neurons were designed. Nickel and Stiefelhagen applied this method to estimate head orientations, which is an additional feature used in ARMAR

[7] Face, human and emotion can be unified as a different categories of objects.

III's pointing gesture detection. For the input of Neural Networks, intensity images and the corresponding disparity images were used. The number of hidden neurons was empirically determined to be 60 [46].

*E. Boosting* Boost methods are employed to generate a strong classifier by using a series of weak classifiers. AdaBoost is a representative method which has also been widely used with Haar-like features for face detection [32]. Due to its fast speed and high detection rate, this method has been widely used in many face detection systems, especially for real-time applications. A representative example is that it has been applied in OpenCV, a famous library of programming functions for real-time computer vision, to detect frontal face. Moreover, some extended methods have been developed in recent years. For example, Spexard *et al.* [205] proposed an improved method in BARTHOC to detect faces with rotation angles including 20°, 40°, 60°, and 80°. In their study, the idea of classification pyramid was added in Viola and Jones's work. In addition to the classification types, speed performance is also improved. Based on the results of face detection, the authors also performed face recognition by using the Eigenfaces method [206] to matching the detected face and stored faces.

Besides AdaBoost, there is another boosting method called Gentle-Boost which represents modified version of the real AdaBoost method [207]. The main difference between these two methods is in the shape of the loss function being used. Based on large features extracted by spatio-temporal box filters, Gentle-Boost was used to construct a strong classifier and has been used in RUBI to recognize children's emotions [37]. Gentle-Boost is designed to estimate sequential maximum likelihood and select features. The proposed methods were used in Berlin dataset, Orator dataset, and their own dataset collected from Early Childhood Education Center at UCSD. Six basic emotion states plus neutral were recognized in the Berlin dataset; seven classes of emotions including agitation, anger, confidence, happiness, leadership, pleasantness, and strength were recognized on the Orator dataset; three categories of sound were recognized on their own dataset.

*F. Hidden Markov Model* Hidden Markov Model (HMM) is a statistical learning technique and has been widely used in speech recognition, emotion recognition, and gesture recognition. For example, a Continuous Density Hidden Markov Model (CDHMM) method was applied in speech recognition based on phoneme. To train the phoneme models, standard European and American databases as well as a specific database with the words were employed [9]. In RoboX, the specific database consists of Yes/No with French, German, Italian, and English. To improve the recognition performance, spatial filtering, dereverberation, and noise cancelation were implemented. In addition to speech recognition, HMM was also applied to recognize pointing gestures. In ARMAR III [46], three dedicated HMM were employed to model hand gestures' phases, including preparation, peak, and retraction. While for the models' inputs, the features representing the positions of the pointing hand and head written in cylindrical coordinate were used. Moreover, to improve the recognition performance, head orientation was deployed as an additional information. In ARMAR III, HMM was also employed to classify sound events in kitchen environments [44]. For the employed HMM, ergodic HMMs with two, three, and four states were tested on one-frame, three-frame, five-frame, seven-frame, and nine-frame with ICA features, respectively. Experimental results have shown that ergodic trisate HMMs with seven frames achieved the lowest error. Another example is HRP-2W service robot who can demonstrate four kinds of motions including walking, squat, picking up, and Cossack dancing by observing and learning human's motions [208]. Mimesis model was proposed to implement the recognition of human's motion and generation of robots' motion by using motion elements and proto-symbols. To acquire motion elements, a hybrid HMMs integrating continuous HMMs (CHMMs) and discrete HMMs (DHMMs) was developed. CHMMs were used in motion elements acquisition phase, and DHMMs were applied in motion recognition. For the designed hybrid HMMs, the input is observed behavior sequences of human, the output is motion elements the robot required, and the used parameter of HMMs is proto-symbols. In the application, each proto-symbol has each corresponding motion.

*G. Gaussian Mixture Models* Gaussian Mixture Models (GMMs) is another statistical model for clustering and classification, which has also been used to sound event classification in ARMAR III [44]. The number of Gaussian models per state can be determined by a Bayesian information criterion. In Kim and colleagues' work [179], GMMs with Bayesian information criterion was deployed to recognize speaker-independent affective states. For different emotion states, the number of Gaussian components is usually different. The Korean affective statesal speech (KES) database was applied for training.

*H. Others* Besides the aforementioned machine learning techniques, fuzzy model, Naive Bayes classifier, and Support vector machine (SVM) have also been employed for recognition. For example, MEXI [51] used fuzzy model [209] to recognize emotions which include happiness, sadness, anger, fear, and neutral. Carotenuto *et al.* [191] applied a Fuzzy model to classify objects into bar, point object, and round object. BARTHOC Jr [173] utilized a Naive Bayes classifier to recognize emotion states including fear, anger, joy, boredom, sadness, and disgust on the Berlin dataset. Altun and Polat [210] used SVM to recognize speech-based emotion states based on Berlin Affective statesal Speech Database-EmoDB.

### 3.3.2 Object Tracking

As we described before, if a robot wants to follow a user's face by the head and neck or a walking person by using its mobile base, object tracking is naturally encountered in social robots. Usually, object tracking can be implemented based on visual-based, range sensors-based, and multimodal-based signals. For visual-based tracking approaches, there are generally two categories including deterministic-based and stochastic-based [211]. For the first category, Gradient descent and mean shift are two popular methods, and Kalman filter and Particle filter are the other two widely used methods for the second category. For range sensors-based tracking methods, Kalman filter and Particle filter are usually applied. Particle filter has attracted increasing attention from social robot researchers due to its high accuracy, strong robustness, and good flexibility in tracking moving objects. Hence, we focus on Particle filter here, especially on its initialization since it is strongly related to the extracted features introduced above.

Particle filter is a sequential important technique which uses a set of weighted samples to approximate the probability density function of a nonlinear system state [211]. For example, Kanda *et al.* [35] used Particle filters in Robovie to track people's trajectories in a shopping arcade. For the initial position of the people to be tracked, frame difference is applied for foreground segmentation. Similarly, Wu *et al.* [212] applied frame difference as detected target, in which color-spatial information was also combined. Since color distributions are insensitive to partial occlusion, invariant to rotation and scale, and computationally efficient, Nummiaro *et al.* [213] also integrated a color distribution into Particle filter. In addition to frame difference and color-related information, other features representing the objects were employed. In Zhai and colleagues' work [211], edges representing human's heads were the measurement cue since they are robust to varying illumination and pose. Kwon *et al.* [214] employed an appearance-adaptive model in Particle filter to handle the occlusion problem. Muñoz-Salinas *et al.* [215] combined depth, color, and gradient information in Particle filter to track multiple persons with different actions such as jumping, running, and shaking hands.

### 3.3.3 Object Segmentation

It is well-known that objects are difficult to be segmented when they have similar texture and color with the environment. However, if objects are moving, they can be easily detected, and the motion of objects can be obtained through a robot's arms which can exert certain force on objects. Based on this observation, Paul and colleagues developed a mechanism and applied it to Cog [12]. The designed mechanism visually probed objects' connectivity and physical extent by adopting reaching action that needs no prior knowledge of objects. The key problem in the designed mechanism is to locate a robot's arms visually. Two methods such as optic flow and image difference were presented to localize the robot's arm from its motion. For the object to be segmented, it can be detected by finding the relation between the object's movement and the robot's end-effector's impact time and spacial position. However, how a robot knows whether the arm collides with an object, how long the exploring procedure, and how about the arms' moving performance still remain unsolved in this area.

### 3.3.4 Speaker Localization

Speaker localization is commonly used in robot's attention system to direct the robot's attention and reduce searching regions in face and human detection. For this semantic understanding task, the raw audio data are collected by microphones, and the cross-power spectrum phase (CSP) is used to compute temporal shift between two audio signals. According to time delay of arrival (TDOA) from temporal shift, the positions of the speaker corresponding to microphones' positions can be obtained. This approach has been used in BARTHOC [205] and Fritz [64] to localize speakers. Shibata and colleagues applied two audio features including ITD and IID to their pet robot. They trained a recurrent neural network to localize sound source by using audio and visual features where visual features such as the motion of humans were training references to generate error signals [18, 181]. The key is to build up a mapping function from sound features including ITD, ILD, and spectral notches for sound source localization in the robot's head spherical coordinates. In the mapping procedure, a linear regression model was built to estimate parameters. Moreover, a Broyden update rule was used for online estimation. For the online positions of sound sources, they can be achieved by using a face detection method.

### 3.3.5 Discussion

In this section, we have introduced the widely used methods for social robots' semantic understanding tasks, including object recognition, object tracking, object segmentation, and speaker localization. The reviewed methods are template matching, clustering, nearest neighbor, Neural Network, Boosting, Hidden Markov Model, and others. Among these methods, the first three are simplest. However, they only demonstrate good performance for some simple applications. For clustering method, it usually works together with template matching and nearest neighbor methods.

For other methods like Neural Network, Boosting, and Hidden Markov Model, they usually require a training process which heavily affects the final recognition results. The advantage of these methods is that they have better generalization ability. Face detection is a good example. First, the face detector will be trained by a large number of face and non-face samples. After training, it can detect face from new images with a high recognition accuracy. However, how to choose suitable training parameters, design the training set are very challenging. Generally speaking, different training parameters are suitable for different applications. If the designed perception system is applied in different scenarios, the parameters of the used method need to be reset, the system will be partially autonomous. Moreover, if there is a big variance between the training and testing data, the performance of these methods will heavily drop.

Besides the above introduced semantic understanding tasks, there are some other ways that can be used to interact with robots, such as a PDA device, a touchscreen, or just buttons. These interaction forms are called "explicit input" by Zhang *et al.* [216] since they are directly operated by user's hand to help robots to accurately implement the input tasks in a simple and intuitive interaction manner. Button was designed in RoboX as a complement for other modalities [9]. Therefore, with respect to language selection, questions responding, and exhibit controlling, if robots cannot give any correct responses, users can accomplish them through button operations. A PDA was applied in robots GRACE [27] and Cero [68]. By using it, users could decide where the robots go. Furthermore, GRACE used a touchscreen to accomplish the same tasks [28].

For semantic understanding tasks in social robots' perception, many methods can be used for different robots even for the same application, and different combinations between features and methods usually demonstrate different performance. For example, Serrano *et al.* [217] reviewed Gabor-based face recognition methods and comparison studies have shown that different recognition approaches with the same feature lead to different recognition accuracies. Devillers *et al.* [171] surveyed features for speech-based emotion recognition with different machine learning techniques. Similarly, Whitehill *et al.* [218] observed that the combinations of Gabor-based features and SVM, and Haar-like features and Gentleboost demonstrate good performance in a practical smile detection system. Since the results of semantic understanding tasks directly affect the efficiency and efficacy of HRI, the used semantic understanding methods and features should be appropriately selected in real applications.

## 3.4 Empirical Comparisons

### *3.4.1 Data Sets*

To evaluate the performance of HRI in social robots, one or more datasets are usually required. Generally speaking, most existing experimental results were conducted on several small datasets, and large-scale benchmark datasets to evaluate the HRI performance of social robots are still required. It is possibly due to the fact that different social robots have different semantic understanding tasks, and the data used in different applications are generally different. However, it is desirable to build up several benchmark datasets for social robots research, which can be a platform for comparing the methods used in different social robots. Hence, several popular data sets used in social robotics are analyzed and compared in light of our analysis of HRI perception methods.

*(1) Audio-Based Dataset*

*(A) Kismet Emotion Dataset*

This dataset contains 5-emotion audio signals collected from two females frequently interacting with Kismet by using a wireless microphone. The expressed affective states include approval, attentional bid, prohibition, soothing, and neutral. Each audio signal is processed to 16-bit signal channel with a frequency of 8 kHz. Finally, 726 labeled samples are selected in the whole dataset [22].

*(B) Berlin Audio Dataset*

The dataset includes 7-emotion audio signals from 10 German actors (5 female and 5 male). The expressed affective states are anger, boredom, disgust, fear, joy, sadness, and neutral. Five long utterances and five short utterances are recorded for each speaker when presenting the emotion. Speech samples are correctly classified by 80 % of human labelers. There are 493 samples, in which 286 and 207 are from the female and male respectively. Each sample has the length of 2–8 s [219].

*C. Orator Dataset*

The dataset consists of 7-emotion audio signals from 13 actors and 14 non-actors in Germany. The expressed affective states are agitation, anger, confidence, happiness, leadership, pleasantness, and strength. There are 150 6-second audio samples that are labeled by 20 labelers in this dataset [220].

*D. RUBI Emotion Dataset*

The RUBI emotion dataset contains recordings from the Early Childhood Education Center (ECEC) at UCSD in one day. The recorded data are coded into three categories: crying, playing/singing, and the background. Each audio signal is non-overlapping. There are 79 crying samples, 72 playing/singing samples, and 151 background samples in this dataset [37].

**Table 6** Performance comparison of some high-level semantic understanding tasks obtained by several representative social robots

| Social Robot | Semantic understanding task | Accuracy | Year |
|---|---|---|---|
| Reckham | face recognition [33] | high | 2007 |
| RUBI | facial action recognition [38] | middle | 2006 |
| | auditory mood detection [37] | middle | 2008 |
| ARMARIII | face verification [42] | low | 2005 |
| | speaker tracking [43] | middle | 2005 |
| | sound event classification [44] | middle | 2005 |
| | head pose estimation [45] | middle | 2007 |
| | gestures recognition [46] | low | 2007 |
| PaPero | speech recognition [49] | low | 2006 |
| MEXI | emotion recognition [53] | middle | 2005 |
| ROMAN | emotion recognition [57] | low | 2008 |
| BARTHOC | emotion recognition [59] | middle | 2005 |
| BIRON | human tracking [62] | high | 2003 |

*E. ARMAR III Sound Dataset*

This dataset consists of audio data collected from four kitchens by using a Sony ECM-719 stereo microphone and a Sony MZ-NH700 High-Minidisc recorder at 44.1 kHz. There are roughly 6000 samples included and manually labeled into 21 categories including boiling, bread cutter, cutting vegetables, door, door bell, egg time ring, footsteps, lighter, match, microwave beep, oven switch, oven timer, over boiling, pan stove, pan sizzling, telephone, speech, stove error, toaster, water, and others [44].

*(2) Visual-Based Dataset*

*A. ARMAR III Gesture Dataset*

The dataset contains 129 pointing gestures collected from 12 subjects in an indoor environment with 8 different pointing targets. Each subject is asked to point one of the marked objects in the field of view of a camera that represents a household robot. Then, the pointing gesture is recorded and added into the dataset with the manual label [46].

*B. ARMAR III Audio-Visual Dataset*

The dataset includes lecture data recorded by 4 fixed cameras and 3 T-shaped microphone arrays (4 microphones in each array) in University of Karlsruhe. The cameras were mounted on the room corners at a height of 2.7 m, and the microphones were placed on the walls excluding the wall behind the lecturer. By using the presented equipment for data collection, a recording usually has the length of 45 min where images were captured at a resolution of 640 × 480 pixels and a frame rate of 15 fps, and sounds were acquired at a sample rate of 44.1 kHz and a resolution of 24 bit [43].

*C. iCub Audio-Visual Dataset*

The dataset contains 184 videos showing 24 different parental couples whose infants have ages from 8 to 30 months. The cup stacking, in which parents demonstrated how the cups were stacked; the wooden bricks, in which tasks parents were instructed to put a block on a pole (altogether three blocks were put); the bell, which rang after pressing the red button; and the salt shaker, which was filled with salt, with the parents demonstrating how to shake the salt on a the blue tray. For the collected videos, images were captured at the resolution of 720 × 576 pixels and a frame rate of 25 fps, and mono sounds were acquired at a sample rate of 44.1 kHz [16].

*D. CK Facial Expression Dataset*

Cohn-Kanade (CK) facial expression database is a widely used visual dataset to evaluate different facial expression recognition methods [221]. It 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 were instructed to perform a series of 23 facial displays, six of which are prototypic emotions mentioned above.

*(3) Tactile-Based Dataset*

Huggable tactile dataset [50] contains tactile signals recoded from the top and left regions of the robot's arm section at a baud rate of 57600. 200 data subsets were included and divided into 16 affective touch interaction types like tickle (softly, fingers only), tickle (hard, fingers only), poking (softly), poking (hard), scratching (one finger softly), scratching (one finger hard), slapping (fingers only softly), slapping (fingers and palm softly), slapping (fingers only hard), Petting (softly), Petting (hard), patting (softly), patting (hard), rubbing, squeezing, and contact.

### *3.4.2 Comparisons of Some Published Results*

In this subsection, we present some representative semantic understanding results that are contained in several social robots literature. Table 6 lists the corresponding semantic understanding tasks and results published.[8] It should be noted that the results cannot be compared directly as the data sets used in different robots are generally different. Moreover, the captured features and semantic understanding tasks are usually different for different robots. We can see from this table that the performance obtained is generally not very encouraging, which shows that there are still some room for researchers to improve in this area.

## 4 Summary

A social robot's perception system is to help the robot understand well the surrounding environment. It can use visual, audio, tactile and laser reading signals to interact with humans. In addition, external equipment such as keyboard, mouse, and touch screen can also be used in HRI. Having acquired these signals, how to process them to obtain useful information is important for robots' perception systems. In this paper, we have reviewed state-of-the-art perception methods in HRI of social robots. We first presented some representative and well-known social robots and then reviewed existing perception methods from three aspects: feature extraction, dimensionality reduction and semantic understanding. For feature extraction, we presented four widely used signals including visual-based, audio-based, tactile-based and range sensors-based. For dimensionality reduction, representative methods including PCA, LDA, and LPP are reviewed. And for the semantic understanding, we reviewed state-of-the-art techniques for several typical applications such as object recognition, object tracking, object segmentation, and speaker localization.

Different from the pure theoretic study, social robots' perception tasks require practicality, which indicates besides high recognition accuracy, the requirements such as low computational cost and autonomy are also very important. For real-time applications, it is desirable to employ the perception methods which have low computational cost, such that the robot can quickly give a response to humans. Another important factor is autonomy that directly determines whether the developed social robots can be commercially available. This is because the consumers of the robots are not experts in the fields, they may not be willing to learn how to adjust the complex parameters of methods to adapt to the working conditions. They hope that the products can automatically adjust and conveniently operate. Therefore, even if the developed social robots cannot achieve fully-autonomy, a higher autonomy should be satisfied. To the best of our knowledge, the reviewed perception methods listed in Tables 1 and 2 for social robots can work in real time and have high autonomy.

The final objective for social robots is to make them work in real environments. While many efforts have been made, there are still some challenging issues to be addressed due to the complex working environments. For feature extraction:

(1) Different features demonstrate different performance in different applications. Therefore, how to efficiently choose the features and how many features are needed for feature fusion in a specified application remain unsolved.
(2) Due to the big difference between the laboratory and real environments, how to apply the methods developed in labs to real-world applications is still a challenging problem. In other words, more robust feature representation and extraction methods are still required.
(3) To improve the performance of semantic understanding tasks, multimodal feature fusion is a popular choice and has been widely used. For instance, Zhang *et al.* [222] proposed local Gabor binary pattern histogram sequence (LGBPHS) to represent face images, in which Gabor magnitude and LBP were combined to demonstrate better performance on the FERET face database. However, how to effectively combine them is still an open problem in this area.

For semantic understanding:

(1) Different semantic understanding methods demonstrate different performance. How to properly combine a semantic understanding method with the extracted features to achieve high performance needs to be further investigated.
(2) Similar to feature extraction, how to apply a semantic understanding method developed in labs to the real environments is still a bottleneck.
(3) Many researchers working on the perception tasks of social robots have achieved encouraging performance. How to improve them to obtain less computational cost and high autonomy needs to be further studied.

To address the above problems, we propose some potential solutions/directions in the following:

(1) For semantic understanding methods, most of them require a training phase first. We have reviewed several commonly used social robot datasets in Sect. 3. However, most of them are collected under controlled conditions such as Orator and CK datasets. There will be a big variance between them and the one collected under spontaneous situations. Undoubtedly, when the developed perception system is used in real environments, the recognition results are heavily affected. Therefore, to ameliorate the training procedure, training datasets

[8] Please note that some other results are not included here. In this table, only the representative social robots listed in Table 1 are selected.

should be acquired from real environments. After training by using these datasets, the final recognition results may be better.

(2) Designing new enhanced methods for feature extraction and semantic understanding is another direction. For example, if the developed features can represent the nature of recognized object under complex environments, it is easy to obtain good results even if the nearest neighbor method is employed. However, this is a very challenging procedure, and it is desirable to collect a large number of real data to improve the performance of existing feature extraction and semantic understanding methods.

(3) The evaluation criteria for perception system in social robots is slightly different from pure research. It is better to achieve high accuracies of perception systems, however, the perception systems serve applications of social robots, and if the robot can give a reasonable response during HRI, humans will not care if its perception system can obtain a very high accuracy. Therefore, in addition to improving the algorithms used in social robots' perception system, designing a reasonable action response of social robots to address the shortcomings of the perception systems is a possible solution.

## References

1. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166
2. Breazeal C (2002) Designing sociable robots. MIT Press, Cambridge
3. Bartneck C, Forlizzi J (2004) A design-centred framework for social human–robot interaction. In: IEEE international workshop on robot and human interactive communication, pp 591–594
4. Hegel F, Muhl C, Wrede B, Martina H-F, Sagerer G (2009) Understanding social robots. In: International conference on advance in computer–human interactions, pp 169–174
5. Social robot, accessed 5 November, 2011 [Online]. Available from: http://en.wikipedia.org/wiki/Social_robot
6. Breazeal C (2003) Toward sociable robots. Robot Auton Syst 42(3–4):167–175
7. Hirose M, Ogawa K (2007) Honda humanoid robots development. Philos Trans R Soc, Math Phys Eng Sci 365:11–19
8. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166
9. Jensen B, Tomatis N, Mayor L, Drygajlo A, Siegwart R (2005) Robots meet humans-interaction in public spaces. IEEE Trans Ind Electron 52(6):1530–1546
10. Jones C, Deeming A (2008) Affective human–robotic interaction. In: Lecture Notes in Computer Science, vol 4868. Springer, Berlin, pp 175–185
11. The FG-NET Aging Database, accessed 25 February, 2008 [Online]. Available from: http://www.fgnet.rsunit.com/
12. Fitzpatrick PM, Metta G (2002) Towards manipulation-driven vision. In: IEEE international conference on intelligent robots and systems, vol 1, pp 43–48
13. Scassellati B (1998) Eye finding via face detection for a foveated, active vision system. In: National conference on artificial intelligence, pp 969–976
14. Tikhanoff V, Cangelosi A, Fitzpatrick P, Metta G, Natale L, Nori F (2008) An open-source simulator for cognitive robotics research: the prototype of the iCub humanoid robot simulator. In: Performance metrics for intelligent systems (PerMIS) workshop, pp 57–61
15. Sandini G, Metta G, Vernon D (2007) The iCub cognitive humanoid robot: an open-system research platform for enactive cognition. In: Lecture notes in computer science, vol 4850. Springer, Berlin, pp 358–369
16. Rolf M, Hanheide M, Rohlfing KJ (2009) Attention via synchrony: making use of multimodal cues in social learning. IEEE Trans Auton Mental Dev 1(1):55–67
17. Figueira D, Lopes M, Ventura R, Ruesch J (2009) Towards a spatial model for humanoid social robots. In: Lecture notes in computer science. Springer, Berlin, pp 287–298
18. Hornstein J, Lopes M, Santos-Victor J, Lacerda F (2006) Sound localization for humanoid robots—building audio-motor maps based on the HRTF. In: International conference on intelligent robots and systems, pp 1170–1176
19. Breazeal C (2003) Emotion and sociable humanoid robots. Int J Hum-Comput Stud 59(1–2):119–155
20. Breazeal C, Edsinger A, Fitzpatrick P, Scassellati B (2001) Active vision for sociable robots. IEEE Trans Syst Man Cybern, Part A, Syst Hum 31(5):443–453
21. Aryananda L (2002) Recognizing and remembering individuals: online and unsupervised face recognition for humanoid robot. In: IEEE international conference on intelligent robots and systems, vol 2, pp 1202–1207
22. Breazeal C, Aryananda L (2002) Recognition of affective communicative intent in robot-directed speech. Auton Robots 12(1):83–104
23. Ge S, Wang C, Hang C (2008) Facial expression imitation in human robot interaction. In: IEEE international symposium on robot and human interactive communication, pp 213–218
24. Barciela G, Paz E, López J, Sanz R, Perez D (2008) Building a robot head: design and control issues. In: IEEE international symposium on robot and human interactive communication, pp 213–218
25. Breazeal C, Kidd CD, Thomaz AL, Hoffman G, Berlin M (2005) Effects of nonverbal communication on efficiency and robustness in human–robot teamwork. In: International conference on intelligent robots and systems, pp 383–388
26. Feil-Seifer D, Matarić MJ (2005) Defining socially assistive robotics. In: International conference on rehabilitation robotics, pp 465–468
27. Simmons R, Goldberg D, Goode A, Montemerlo M, Roy N, Sellner B, Urmson C, Maxwell B (2003) GRACE: an autonomous robot for the AAAI robot challenge. AI Mag 24(2):51–72
28. Michalowski MP, Šabanović S, Disalvo C, Busquets D, Hiatt LM, Melchior NA, Simmons R (2007) Socially distributed perception: GRACE plays social tag at AAAI 2005. Auton Robots 22(4):385–397
29. Clodic A, Fleury S, Alami R, Herrb M, Chatila R (2005) Supervision and interaction. In: International conference on advanced robotics, pp 725–732
30. Jensen B, Philippsen R, Siegwart R (2003) Narrative situation assessment for human–robot interaction. In: IEEE international conference on robotics and automation, vol 1, pp 1503–1508
31. Jensen B, Froidevaux G, Greppin X, Lorotte A, Mayor L, Meisser M, Ramel G, Siegwart R (2003) Multi-robot human-interaction and visitor flow management. In: IEEE international conference on robotics and automation, pp 2388–2393

32. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: IEEE computer society conference on computer vision and pattern recognition, vol 1, pp I511–I518
33. Germa T, Lerasle F, Danès P, Brèthes L (2007) Human/robot visual interaction for a tour-guide robot. In: IEEE international conference on intelligent robots and systems, pp 3448–3453
34. Hasanuzzaman Md, Zhang T, Ampornaramveth V, Gotoda H, Shirai Y, Ueno H (2007) Adaptive visual gesture recognition for human–robot interaction using a knowledge-based software platform. Robot Auton Syst 55(8):643–657
35. Kanda T, Glas DF, Shiomi M (2009) Abstracting people's trajectories for social robots to proactively approach customers. IEEE Trans Robot 25(6):1382–1396
36. Movellan JR, Tanaka F, Fasel IR, Taylor C, Ruvolo P, Eckhardt M (2007) The RUBI project: a progress report. In: ACM/IEEE conference on human–robot interaction—robot as team member, pp 333–339
37. Ruvolo P, Fasel I, Movellan J (2008) Auditory mood detection for social and educational robots. In: IEEE international conference on robotics and automation, pp 3551–3556
38. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2006) Fully automatic facial action recognition in spontaneous behavior. In: International conference on automatic face and gesture recognition, pp 223–230
39. Christensen HI (2003) Intelligent home appliances. In: Springer tracts in advanced robotics. Springer, Berlin, pp 319–330
40. Lohse M, Hegel F, Wrede B (2008) Domestic applications for social robots-an online survey on the influence of appearance and capabilities. J Phys Agents 2(2):21–32
41. Asfour T, Regenstein K, Azad P, Schroder O, Bierbaum A, Vahrenkamp N, Dillmann R (2006) ARMAR-III: an integrated humanoid platform for sensory-motor control. In: International conference on humanoid robots, pp 169–175
42. Ekenel HK, Stiefelhagen R (2005) A generic face representation approach for local appearance based face verification. In: IEEE computer society conference on computer vision and pattern recognition workshops, vol 03, p 155
43. Nickel K, Gehrig T, Stiefelhagen R, McDonough J (2005) A joint particle filter for audio-visual speaker tracking. In: International conference on multimodal interfaces, pp 61–68
44. Kraft F, Malkin R, Schaaf T, Waibel A (2005) Temporal ICA for classification of acoustic events in a kitchen environment. In: European conference on speech communication and technology, pp 2689–2692
45. Voit M, Nickel K, Stiefelhagen R (2007) Neural network-based head pose estimation and multi-view fusion. In: Lecture notes in computer science, vol 4122. Springer, Berlin, pp 291–298
46. Nickel K, Stiefelhagen R (2007) Visual recognition of pointing gestures for human–robot interaction. Image Vis Comput 25(12):1875–1884
47. Osada J, Ohnaka S, Sato M (2006) The scenario and design process of childcare robot. In: PaPeRo, international conference on advances in computer entertainment technology. Springer, Berlin
48. Sato A, Imaoka H, Suzuki T, Hosoi T (2005) Advances in face detection and recognition technologies. NEC J Adv Technol 2(1):28–34
49. Betkowska A, Shinoda K, Furui S (2007) Robust speech recognition using factorial HMMs for home environments. Eurasip J Adv Signal Process. doi:10.1155/2007/20593
50. Stiehl W, Breazeal C (2005) Affective touch for robotic companions. In: Lecture notes in computer science, vol 3784. Springer, Berlin, pp 747–754
51. Esau N, Kleinjohann L, Kleinjohann B (2006) Emotional communication with the robot head MEXI. In: International conference on control, automation, robotics and vision, pp 1–7
52. Stichling D, Kleinjohann B (2002) Low latency color segmentation on embedded real-time systems. In: IFIP world computer congress—TC10 stream on distributed and parallel embedded systems, vol 219, pp 247–256
53. Austermann A, Esa N, Kleinjohann L, Kleinjohann B (2005) Prosody based emotion recognition for MEXI. In: International conference on intelligent robots and systems, vol 3, pp 1138–1144
54. Esau N, Kleinjohann L, Kleinjohann B (2005) An adaptable fuzzy affective states model for affective states recognition. In: EUSFLAT—LFA, pp 73–78
55. Hirth J, Schmitz N, Berns K (2007) Emotional architecture for the humanoid robot head ROMAN. In: IEEE international conference on robotics and automation, pp 2150–2155
56. Schmitz N, Spranger C, Berns K (2009) 3D audio perception system for humanoid robots. In: International conferences on advances in computer–human interactions, pp 181–186
57. Strupp S, Schmitz N, Berns K (2008) Visual-based emotion detection for natural man–machine interaction. In: Lecture notes in computer science, vol 5243. Springer, Berlin, pp 356–363
58. Hackel M, Schwope S, Fritsch J, Wrede B, Sagerer G (2006) Designing a sociable humanoid robot for interdisciplinary research. Adv Robot 20(11):1219–1235
59. Vogt T, Andreé E (2005) Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: IEEE international conference on multimedia and expo, pp 474–477
60. Spexard T, Haasch A, Fritsch J, Sagerer G (2006) Human-like person tracking with an anthropomorphic robot. In: IEEE international conference on robotics and automation, pp 1286–1292
61. Haasch A, Hohenner S, Hüwel S, Kleinehagenbrock M, Lang S, Toptsis I, Fink G, Fritsch J, Wrede B, Sagerer G (2004) BIRON-the bielefeld robot companion. In: International workshop on advances in service robotics, pp 27–32
62. Fritsch J, Kleinehagenbrock M, Lang S, Plötz T, Fink GA, Sagerer G (2003) Multi-modal anchoring for human–robot interaction. Robot Auton Syst 43(2–3):133–147
63. Lang S, Kleinehagenbrock M, Hohenner S, Fritsch J, Fink GA, Sagerer G (2003) Providing the basis for human–robot-interaction: a multi-modal attention system for a mobile robot. In: International conference on multimodal interfaces, pp 28–35
64. Bennewitz M, Faber F, Joho D, Behnke S (2007) Fritz—a humanoid communication robot. In: IEEE international conference on robot & human interactive communication, pp 1072–1077
65. Lisetti CL, Brown SM, Alvarez K, Marpaung AH (2004) A social informatics approach to human–robot interaction with a service social robot. IEEE Trans Syst Man Cybern, Part C, Appl Rev 34(2):195–209
66. Brown SM, Lisetti CL, Marpaung AH (2002) Cherry, the little red robot…with a mission…and a personality. In: AAAI fall symposium
67. Marpaung AH, Lisetti CL (2002) Multilevel emotion modeling for autonomous agents. In: AAAI fall symposium—technical report FS-04-05, pp 39–46
68. Kerstin S, Anders G, Helge H (2003) Social and collaborative aspects of interaction with a service robot. Robot Auton Syst 42:223–234
69. Chopra A, Obsniuk M, Jenkin MR (2006) The nomad 200 and the nomad SuperScout: reverse engineered and resurrected. In: Canadian conference on computer and robot vision
70. Kozima H, Michalowski M, Nakagawa C (2009) A playful robot for research, therapy, and entertainment. Int J Soc Robot 1:3–18
71. Wada K, Shibata T (2007) Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. IEEE Trans Robot 23(5):972–980

72. Goris K, Saldien J, Lefeber D (2008) Probo, a testbed for human robot interaction. In: ACM/IEEE international conference on human–robot interaction, pp 253–254
73. Saldien J, Goris K, Vanderborght B, Lefeber D (2008) On the design of an emotional interface for the huggable robot probo. In: The reign of catz and dogz, AISB2008
74. Goris K, Saldien J, Vanderborght B, Lefeber D (2008) The huggable robot probo: design of a robotic head. In: The reign of catz and dogz, AISB2008
75. Poel M, Heylen D, Nijholt A, Meulemans M, Breemen A (2009) Gaze behaviour, believability, likability and the iCat. AI Soc 24:61–73
76. Van Breemen AJN (2004) Animation engine for believable interactive user-interface robots. In: IEEE/RSJ international conference on intelligent robots and systems, vol 3, pp 2873–2878
77. Ronald C, Fujita M, Tsuyoshi T, Rika H (2003) An ethological and emotional basis for human–robot interaction. Robot Auton Syst 42:191–201
78. Oh JH, Hanson D, Kim WS, Han IY, Kim JY, Park IW (2006) Design of android type humanoid robot Albert HUBO. In: IEEE international conference on intelligent robots and systems, pp 1428–1433
79. Miwa H, Itoh K, Matsumoto M, Zecca M, Takanobu H, Roccella S, Carrozza MC, Takanishi A (2004) Effective affective statesal expressions with affective states expression humanoid robot WE-4RII—integration of humanoid robot hand RCH-1. In: International conference on intelligent robots and systems, vol 3, pp 2203–2208
80. Ogura Y, Aikawa H, Shimomura K, Kondo H, Morishima A, Lim HO, Takanishi A (2006) Development of a new humanoid robot WABIAN-2. In: IEEE international conference on robotics and automation, pp 76–81
81. Zecca M, Mizoguch Y, Endo K, Iida F, Kawabata Y, Endo N, Itoh K, Takanishi A (2009) Whole body emotion expressions for expressions for KOBIAN humanoid robot-preliminary experiments with different affective statesal patterns. In: IEEE international workshop on robot and human interactive communication, pp 381–386
82. Salichs MA, Barber R, Khamis AM, Malfaz M, Gorostiza JF, Pacheco R, Rivas R, García D (2006) Maggie: a robotic platform for human–robot social interaction. In: IEEE conference on robotics, automation and mechatronics
83. Gorostiza J, Barber R, Khamis A, Pacheco M, Rivas R, Corrales A, Delgado E, Salichs M (2006) Multimodal human–robot interaction framework for a personal robot. In: International symposium on robot and human interactive communication, pp 39–44
84. Kormushev P, Nenchev DN, Calinon S, Caldwell DG (2011) Upper-body kinesthetic teaching of a free-standing humanoid robot. In: International conference on robotics and automation, pp 3970–3975
85. Ishida T, Kuroki Y, Yamaguchi J (2003) Development of mechanical system for a small biped entertainment robot. In: International workshop on robot and human interactive communication, pp 297–302
86. Park IW, Kim JY, Lee J, Oh JH (2005) Mechanical design of humanoid robot platform KHR-3 (KAIST humanoid robot—3: HUBO). In: International conference on humanoid robots, pp 321–326
87. Okada K, Ogura T, Haneda A, Kousaka D, Nakai H, Inaba M, Inoue H (2004) Integrated system software for HRP2 humanoid. In: International conference on robotics and automation, vol 4, pp 3207–3212
88. Cousins S (2010) ROS on the PR2. IEEE Robot Autom Mag 17(3):23–25
89. Bischoff R, Huggenberger U, Prassler E (2011) KUKA youBot-a mobile manipulator for research and education. In: International conference on robotics and automation, pp 1–4
90. Goodrich MA, Schultz AC (2007) Human–robot interaction: a survey. Found Trends Hum-Comput Interact 1(3):203–275
91. Castleman KR (1996) Digital image processing. Prentice Hall, New York
92. Kinect, Accessed 9 December, 2011 [Online] Available from: http://en.wikipedia.org/wiki/Kinect
93. Bumblebee2, Accessed 2010 [Online] Available from: http://www.ptgrey.com/products/stereo.asp
94. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
95. Darrell T, Gordon GM, Harville M, Woodfill J (2000) Integrated person tracking using stereo, color, and pattern detection. Int J Comput Vis 37(2):175–185
96. Wang X, Xu H, Wang H, Li H (2008) Robust real-time face detection with skin color detection and the modified census transform. In: IEEE international conference on information and automation, pp 590–595
97. Kakumanu P, Makrogiannis S, Bourbakis N (2007) A survey of skin-color modeling and detection methods. Pattern Recognit 40:1106–1122
98. Ford A, Roberts A (1998) Colour space conversions
99. Ruesch J, Lopes M, Bernardino A, Hörnstein J, Santos-Victor J, Pfeifer R (2008) Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In: IEEE international conference on robotics and automation, pp 962–967
100. Chen J, Tiddeman B (2007) Facial feature detection under various illuminations. In: Lecture notes in computer science, vol 4841. Springer, Berlin, pp 498–508
101. Zabih R, Woodfill J (1994) Non-parametric local transforms for computing visual correspondence. In: European conference on computer vision, pp 151–158
102. Song M, Tao D, Liu Z, Li X, Zhou M (2009) Image ratio features for facial expression recognition application. IEEE Trans Syst Man Cyber Part B Cyber. doi:10.1109/TSMCB.2009.2029076
103. Wang L, He D-C (1990) Texture classification using texture spectrum. Pattern Recognit 23(8):905–910
104. Ojala T, Pietikäinen M, Harwood D (1996) Texture classification using texture spectrum. Pattern Recognit 29(1):51–59
105. Ojala T, Pietikäine M, Mäenpää T (2002) Multiresolution grayscale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
106. Jin H, Liu Q, Lu H, Tong X (2004) Face detection using improved LBP under Bayesian framework. In: International conference on image and graphics, pp 306–309
107. Ahonen T, Hadid A, Matti P (2006) Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Anal Mach Intell 28(12):2037–2041
108. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816
109. Solar J, Quinteros J (2008) Illumination compensation and normalization in eigenspace-based face recognition: a comparative study of different pre-processing approaches. Pattern Recognit Lett 29:1966–1979
110. Zhao G, Pietikainen M (2009) Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recognit Lett 30(12):1117–1127
111. Zabih R, Woodfill J (1996) A non-parametric approach to visual correspondence. IEEE Trans Pattern Anal Mach Intell
112. Christian K, Ernst A (2006) Face detection and tracking in video sequences using the modified census transformation. Image Vis Comput 24:564–572
113. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110

114. Lowe DG (1999) Object recognition from local scale-invariant features. In: International conference on computer vision, pp 1150–1157
115. Tian YL, Kanade T, Conn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115
116. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. Int J Comput Vis 1(4):321–331
117. Illingworth J, Kittler J (1987) The adaptive hough transform. IEEE Trans Pattern Anal Mach Intell 9(5):690–698
118. Levi K, Weiss Y (2004) Learning object detection from a small number of examples: the importance of good features. In: IEEE computer society conference on computer vision and pattern recognition, vol 2, pp II53–II60
119. Raman M, Himanshu A (2009) Study and comparison of various image edge detection techniques. Int J Image Process 3(1):1–12
120. Horn BKP, Schunck BG (1981) Determining optical flow. Artif Intell 17(1–3):185–203
121. Brox T, Bruhn A, Papenberg N, Weickert J (2004) High accuracy optical flow using a theory for warping. In: Lecture notes in computer science, vol 3024. Springer, Berlin, pp 25–36
122. Bab-Hadiashar A, Suter D (1998) Robust optic flow computation. Int J Comput Vis 29(1):59–77
123. Iida F (2003) Biologically inspired visual odometer for navigation of a flying robot. Robot Auton Syst 44:201–208
124. Cédras C, Shah M (1995) Motion-based recognition: a survey. Image Vis Comput 13(2):129–155
125. Moeslund T, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Underst 81:231–268
126. Wang J, Singh S (2003) Video analysis of human dynamics: a survey. Real-Time Imaging 9:321–346
127. Lu J, Zhang E (2007) Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fusion. Pattern Recognit Lett 28(16):2401–2411
128. Lu J, Tan Y-P (2010) Uncorrelated discriminant nearest feature line analysis for face recognition. IEEE Signal Process Lett 17(2):185–188
129. Lu J, Tan Y-P (2010) Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. Pattern Recognit Lett 31(5):382–393
130. Lu J, Tan Y-P (2010) Gait-based human age estimation. IEEE Trans Inf Forensics Secur 5(4):761–770
131. Lu J (2010) Enhanced locality sensitive discriminant analysis for image recognition. Electron Lett 46(3):217–218
132. Lu J, Tan Y-P (2010) A doubly weighted approach for appearance-based subspace learning methods. IEEE Trans Inf Forensics Secur 5(1):71–81
133. Lu J, Tan Y-P (2010) Regularized locality preserving projections and its extensions for face recognition. IEEE Trans Syst Man Cybern, Part B, Cybern 40(2):958–963
134. Lu J, Tan Y-P (2010) Cost-sensitive subspace learning for face recognition. In: IEEE international conference on computer vision and pattern recognition, pp 2661–2666
135. Lu J, Tan Y-P (2011) Nearest feature space analysis for classification. IEEE Signal Process Lett 18(1):55–58
136. Liu N, Lu J, Tan Y-P (2011) Joint subspace learning for view-invariant gait recognition. IEEE Signal Process Lett 18(7):431–434
137. Lu J, Zhou X, Tan Y-P, Shang Y, Zhou J (2012) Cost-sensitive semi-supervised discriminant analysis for face recognition. IEEE Trans Inf Forensics Secur 7(3):944–953
138. Lu J, Tan Y-P (2013) Cost-sensitive subspace analysis and extensions for face recognition. IEEE Trans Inf Forensics Secur 7(3):510–519
139. Lu J, Tan Y-P, Wang G (2013) Discriminative multimanifold analysis for face recognition from a single training sample per person. IEEE Trans Pattern Anal Mach Intell 35(1):39–51
140. Lu J, Zhang E, Kang X, Xue Y, Chen Y (2006) Palmprint recognition using wavelet decomposition and 2D principal component analysis. In: International conference on communications, circuits and systems proceedings, pp 2133–2136
141. Lu J, Zhao Y, Xue Y, Hu J (2008) Palmprint recognition via locality preserving projections and extreme learning machine neural network. In: International conference on signal processing, pp 2096–2099
142. Zhang E, Lu J, Duan G (2005) Gait recognition via independent component analysis based on support vector machine and neural network. In: International conference on natural computation, pp 640–649
143. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
144. Belhumenur PN, Hepanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherface: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
145. He X, Yan S, Hu Y, Niyogi P, Zhang HJ (2005) Face recognition using Laplacian faces. IEEE Trans Pattern Anal Mach Intell 27(3):328–340
146. Dabbaghchian SP, Ghaemmaghami M, Aghagolzadeh A (2010) Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology. Pattern Recognit 43:1431–1440
147. Stiefelhagen R, Ekenel HK, Fügen C, Gieselmann P, Holzapfel H, Kraft F, Nickel K, Waibel A (2007) Enabling multimodal human–robot interaction for the Karlsruhe humanoid robot. IEEE Trans Robot 23(5):840–851
148. Liu C, Wechsler H (2002) Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans Image Process 11(4):467–476
149. Lu J, Zhao Y, Hu J (2009) Enhanced Gabor-based region covariance matrices for palmprint recognition. Electron Lett 45(17):880–881
150. Tong Y, Liao W, Ji Q (2007) Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans Pattern Anal Mach Intell 29(10):1683–1699
151. Susskind JM, Littlewort G, Bartlett MS (2007) Human and computer recognition of facial expressions of emotion. Neuropsychologia 45(1):152–162
152. Pavani SK, Delgado D, Frangi AF (2010) Haar-like features with optimally weighted rectangles for rapid object detection. Pattern Recognit 43(1):160–172
153. Papageorgiou CP, Oren M, Poggio T (1998) A general framework for object detection. In: IEEE international conference on computer vision, pp 555–562
154. Yang P, Li Q, Metaxas DN (2009) Boosting encoded dynamic features for facial expression recognition. Pattern Recognit Lett 30(2):132–139
155. Lai K, Bo L, Ren X, Fox D (2011) A large-scale hierarchical multi-view RGB-D object dataset. In: IEEE international conference on robotics and automation, pp 1817–1824
156. Martin AF, Robert, CB (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395
157. Benavidez P, Jamshidi M (2011) Mobile robot navigation and target tracking system. In: International conference on system of systems engineering, pp 299–304
158. Johnson A, Hebert M (1999) Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans Pattern Anal Mach Intell 21(5):433–449
159. Rusu RB, Blodow N, Beetz M (2009) Fast point feature histograms (FPFH) for 3D registration. In: IEEE international conference on robotics and automation, pp 3212–3217

160. Bo L, Ren X, Fox D (2011) Depth kernel descriptors for object recognition. In: International conference on intelligent robots and systems, pp 821–826
161. Hartley R, Zisserman A (2000) Multiple view geometry in computer vision. Cambridge University Press, Cambridge, pp 1–12
162. Kim I, Kim D, Cha Y, Lee K, Kuc T (2007) An embodiment of stereo vision system for mobile robot for real-time measuring distance and object tracking. In: International conference on control, automation and systems, pp 1029–1033
163. Li Z, Jarvis R (2009) A multi-modal gesture recognition system in a human–robot interaction scenario. In: International workshop on robotic and sensors environments, pp 41–46
164. Thompson S, Kagami S (2005) Humanoid robot localisation using stereo vision. In: International conference on humanoid robots, pp 19–25
165. Prasad R, Saruwatari H, Shikano K (2004) Robots that can hear, understand and talk. Adv Robot 18(5):533–564
166. Sweeney L, Thompson P (1997) Speech perception using realtime phoneme detection: the BeBe system
167. Jaisal PK, Mishra PK (2012) A review of speech pattern recognition: survey. Int J Comput Sci Technol 3(1):709–713
168. Clavel C, Vasilescu I, Devillers L, Richard G, Ehrette T (2008) Fear-type emotion recognition for future audio-based surveillance systems. Speech Commun 50:487–503
169. Vogt T, André E, Johannes W (2008) Automatic recognition of emotion from speech: a review of the literature and recommendations for practical realisation. In: Lecture note in computer science, vol 4868. Springer, Berlin, pp 75–91
170. Hyun K, Kim E, Kwak Y (2007) Emotional feature extraction based on phoneme information for speech emotion recognition. In: IEEE international conference on robot & human interactive communication, pp 802–806
171. Devillers L, Vidrascu L, Lamel L (2005) Challenges in real-life emotion annotation and machine learning based detection. Neural Netw 18:407–422
172. Rong J, Gang L, Chen Y (2008) Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag 45(3):315–328
173. Hegel F, Spexard T, Wrede B, Horstmann G, Vogt T (2006) Playing a different imitation game: interaction with an empathic android robot. In: IEEE-RAS international conference on humanoid robots, pp 56–61
174. Morrison D, Wang R, Silva L (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Commun 49:98–112
175. Markel JD (1972) The SIFT algorithm for fundamental frequency estimation. IEEE Trans Audio Electroacoust AU-20(5):367–377
176. Wang C, Seneff S (2000) Robust pitch tracking for prosodic modeling in telephone speech. In: IEEE international conference on acoustics, speech and signal processing, vol 3, pp 1343–1346
177. Ahmadi S, Spanias AS (1999) Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. IEEE Trans Speech Audio Process 7(3):333–338
178. Xu M, Duan LY, Cai J, Chia, LT, Xu C, Tian Q (2004) CHMM-based audio keyword generation. In: Lecture notes in computer science, vol 3333. Springer, Berlin, pp 566–574
179. Kim E, Hyun K, Kim S, Kwak Y (2009) Improved emotion recognition with a novel speaker-independent feature, IEEE/ASME Trans Mechatron. doi:10.1109/TMECH.2008.2008644
180. Welch P (1967) The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust AU-15:70–73
181. Shibata T, Inoue K, Irie R (1996) Affective statesal robot for intelligent system—artificial affective statesal creature project. In: IEEE international workshop on robot and human communication, pp 466–471
182. Dao D, Sugiyama S (2006) Fabrication and characterization of 4-DOF soft-contact tactile sensor and application to robot fingers. In: International symposium on micro-NanoMechatronics and human science, pp 1–6
183. Tsetserukou D, Kawakami N, Tachi S (2008) An approach to contact force vector determination and its implementation to provide intelligent tactile interaction with environment. In: Lecture notes in computer science, vol 5024. Springer, Berlin, pp 151–156
184. Iwata H, Hoshino H, Morita T, Sugano S (2001) Force detectable surface covers for humanoid robots. In: International conference on advanced intelligent mechatronics, pp 1205–1210
185. Stiehl W, Breazeal C (2006) A sensitive skin for robotic companions featuring temperature, force, and electric field sensors. In: IEEE/RSJ international conference on intelligent robots and systems, pp 1952–1959
186. Stiehl W, Lieberman J, Breazeal C, Basel L, Lalla L, Wolf M (2005) Design of a therapeutic robotic companion for relational, affective touch. In: IEEE international workshop on robots and human interactive communication, pp 408–415
187. Shibata T (2004) Ubiquitous surface tactile sensor. In: IEEE technical exhibition based conference on robotics and automation, pp 5–6
188. Berger DA (1988) On using a tactile sensor for real-time feature extraction. Master's thesis, Carnegie-Mellon University
189. Iwata H, Sugano S (2005) Human–robot-contact-state identification based on tactile recognition. IEEE Trans Ind Electron 52(6):1468–1477
190. Göger D, Gorges N, Wörn H (2009) Tactile sensing for an anthropomorphic robotic hand: hardware and signal processing. In: IEEE international conference on robotics and automation, pp 895–901
191. Carotenuto L, Famularo D, Muraca P, Raiconi G (1997) A fuzzy classifier for tactile sensing. J Intell Robot Syst Theory Appl 20(1):71–86
192. Glas DF, Miyashit T, Ishiguro H, Hagita N (2007) Laser tracking of human body motion using adaptive shape modeling. In: IEEE international conference on intelligent robots and systems, pp 602–608
193. Gockley R, Forlizzi J, Simmons R (2007) Natural person-following behavior for social robots. In: ACM/IEEE international conference on human-robot interaction, pp 17–24
194. Jung B, Sukhatme GS (2009) Real-time motion tracking from a mobile robot. Int J Soc Robot. doi:10.1007/s12369-009-0038
195. Glas DF, Miyashit T, Ishiguro H, Hagita N (2009) Laser-based tracking of human position and orientation using parametric shape modeling. Adv Robot 23:405–428
196. Morales J, Martinez JL, Mandow A, Pequeno-Boter A, Garcia-Cerezo A (2011) Design and development of a fast and precise low-cost 3D laser rangefinder. In: International conference on mechatronics, pp 621–626
197. Scholer F, Behley J, Steinhage V, Schulz D, Cremers AB (2011) Person tracking in three-dimensional laser range data with explicit occlusion adaption. In: International conference on robotics and automation, pp 1297–1303
198. Spinello L, Arras KO, Triebel R, Siegwart R (2010) A layered approach to people detection in 3D range data. In: Proceedings of the national conference on artificial intelligence, vol 3, pp 1625–1630
199. Navarro-Serment LE, Mertz C, Hebert M (2010) Pedestrian detection and tracking using three-dimensional LADAR data. Int J Robot Res 29(12):1516–1528
200. Harrison A, Newman P (2008) High quality 3D laser ranging under general vehicle motion. In: International conference on robotics and automation, pp 7–12

201. Pantic M, Leon J (2003) Toward an affect-sensitive multimodal human–computer interaction. Proc IEEE 91(9):1370–1390
202. Fragopanagos N, Taylor J (2005) Emotion recognition in human–computer interaction. Neural Netw 18(4):389–405
203. Zeng Z, Tu J, Brian M, Huang T (2008) Audio-visual affective expression recognition through multistream fused HMM. IEEE Trans Multimed 10(4):570–577
204. Johnson DO, Agah A (2009) Human robot interaction through semantic integration of multiple modalities, dialog management, and contexts. Int J Soc Robot 1:283–305
205. Spexard T, Hanheide M (2007) Gerhard sagerer, human-oriented interaction with an anthropomorphic robot. IEEE Trans Robot 23(5):852–862
206. Turk M, Pentland A (1991) Eigedces for recognition. J Cogn Neurosci 3(1):71–86
207. Guillaume L, Miroslav R (2009). Directed reading: boosting algorithms
208. Inamura T, Toshima I, Nakamura Y (2003) Acquiring motion elements for bidirectional computation of motion recognition and generation. Exp Robot VIII, 5:372–381
209. Esau N, Kleinjohann L, Kleinjohann B (2005) An adaptable fuzzy affective states model for affective states recognition. In: European society for fuzzy logic and technology, pp 73–78
210. Altun H, Polat G (2009) Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Syst Appl 36(4):8197–8203
211. Zhai Y, Yeary MB, Cheng S, Kehtarnavaz N (2009) An object-tracking algorithm based on multiple-model particle filtering with state partitioning. IEEE Trans Instrum Meas 58(5):1797–1809
212. Wu X, Gong H, Chen P, Zhong Z, Xu Y (2009) Surveillance robot utilizing video and audio information. J Intell Robot Syst 55(4–5):403–421
213. Nummiaro K, Koller-Meier, E, Van Gool, L (2003) An adaptive color-based particle filter. Image Vis Comput 21(1):99–110
214. Kwon HS, Kim, YJ, Lim, MT (2005) Person tracking with a mobile robot using particle filters in complex environment. In: International society for optical engineering, vol 6042. SPIE Press, Bellingham
215. Muñoz-Salinas R, García-Silvente M, Medina Carnicer R (2008) Adaptive multi-modal stereo people tracking without background modelling. J Vis Commun Image Represent 19(2):75–91
216. Tao Z, Biwen Z, Lee L, Kaber D (2008) Service robot anthropomorphism and interface design for emotion in human–robot interaction. In: IEEE conference on automation science and engineering, pp 674–679
217. Serrano A, de Diego IM, Conde C, Cabello E (2009) Recent advances in face biometrics with Gabor wavelets: a review. Pattern Recogn Lett. doi:10.1016/j.patrec.2009.11.002
218. Whitehill J, Littlewort G, Fasel I, Bartlett M, Movellan J (2009) Toward practical smile detection. IEEE Trans Pattern Anal Mach Intell 31(11):2106–2111
219. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: European conference on speech communication and technology, pp 1517–1520
220. Quast H (2001) Automatic recognition of nonverbal speech: an approach to model the perception of para- and extralinguistic vocal communication with neural networks. Master's thesis, University of Göttingen
221. Kanade T, Cohn J, Tian YL (2000) Comprehensive database for facial expression analysis. In: IEEE international conference on face and gesture analysis, pp 46–53
222. Zhang W, Shan S, Gao W, Chen X, Zhang H (2005) Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. In: IEEE international conference on computer vision, vol 1, pp 786–791
223. Goodrich MA, Schultz AC (2007) Human–robot interaction: a survey. Found Trends Hum-Comput Interact 1(3):203–275
224. Drury JL, Scholtz J, Yanco HA (2004) Applying CSCW and HCI techniques to human–robot interaction. In: CHI 2004 workshop on shaping human–robot interaction
225. Moller A, Roalter L, Kranz M (2011) Cognitive objects for human–computer interaction and human–robot interaction. In: HRI2011, 6–9 March, Lausanne, Switzerland
226. Saldien J, Goris K, Vanderborght B, Vanderfaeillie J, Lefeber D (2010) Cognitive Objects Hum-Comput Interact Hum-Robot Interact 2(4):377–389

**Haibin Yan** received the B.Eng. and M.Eng. degrees from the Xi'an University of Technology, Xi'an, China, in 2004 and 2007, and the Ph.D. degree from the National University of Singapore, Singapore, in 2013, all in mechanical engineering. Her research interests include human–robotic interaction, social robotics, and computer vision.

**Marcelo H. Ang Jr.** received the B.Sc. degrees (Cum Laude) in Mechanical Engineering and Industrial Management Engineering from the De La Salle University, Manila, Philippines, in 1981; the M.Sc. degree in Mechanical Engineering from the University of Hawaii at Manoa, Honolulu, Hawaii, in 1985; and the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Rochester, Rochester, New York, in 1986 and 1988, respectively. His work experience includes heading the Technical Training Division of Intel's Assembly and Test Facility in the Philippines, research positions at the East West Center in Hawaii and at the Massachusetts Institute of Technology, and a faculty position as an Assistant Professor of Electrical Engineering at the University of Rochester, New York. In 1989, Dr. Ang joined the Department of Mechanical Engineering of the National University of Singapore, where he is currently an Associate Professor, with a Joint Appointment at the Division of Engineering and Technology Management. He also is the Acting Director of the Advanced Robotics Center, and Dy Director of the Center for Intelligent Products and Manufacturing Systems. His research interests span the areas of robotics, mechatronics, and applications of intelligent systems methodologies. He teaches both at the graduate and undergraduate levels in the following areas: robotics; creativity and innovation, applied electronics and instrumentation; advanced computing; product design and realization. He is also active in consulting work in these areas. In addition to academic and research activities, he is actively involved in the Singapore Robotic Games as its founding chairman.

**Aun Neow Poo** received his B.Eng. degree with first class honors in mechanical engineering from the National University of Singapore in 1968. He then proceeded, on a Ford Foundation Fellowship, to the University of Wisconsin where he obtained his M.Sc. and Ph.D. degrees in 1970 and 1973, respectively. After a year as a research fellow with IBM Thomas Watson Research Centre in New York, he joined the National University of Singapore (NUS) where he is currently professor in the Department of Mechanical Engineering. At NUS, he has held, amongst others, the positions of Head of the Department of Mechanical and Production Engineering, Dean of the Faculty of Engineering, and Director of the Graduate School of Engineering. He is a member of the Senate at NUS and has also been on the Council of both NUS and NTU. His special interest is in automation and mechatronics, a field in which he has worked for more than thirty years, supervised some dozens of doctoral and masters students. He has also been actively involved in organizing numerous international conferences on automation, robotics and control, and served as editor to several international journals. He is a past president and a Fellow of the Institution of Engineers, Singapore, and a Founding Fellow of the Asean

Academy of Engineering and Technology. He has also served on the boards of management of numerous industrial and professional bodies including on Singapore Aerospace Industries, Singapore Institute of Standards and Industrial Research, the Defence Science Organization, the International Federation of Robotics, and GINTIC Institute of Manufacturing Technology.