Interactive Communication Between Human and Robot Using Nonverbal Cues

Salah Al-Darraji[1], Zuhair Zafar[2], Karsten Berns[2], Djordje Urukalo[3], and Aleksandar Rodić[3]

[1] Computer Science Department, University of Basrah, Iraq
aldarraji@uobasrah.edu.iq
[2] Robotics Research Lab., Computer Science Department
University of Kaiserslautern, Kaiserslautern, Germany
{zafar;berns}@cs.uni-kl.de
[3] University of Belgrade, Mihajlo Pupin Institute, Robotics Laboratory, Serbia
{djordje.urukalo;aleksandar.rodic}@pupin.rs

Abstract. Socially interactive robots need the same behaviors and capabilities of people to interact naturally with humans. One of these capabilities is the perception of nonverbal cues. The paper presents a biologically inspired perception system for a social robot. This system is based on the psychological theory of perceptual cycle. It is composed of two main parts: schema and exploration. The schema represents the bottom-up information processing, whereas the exploration represents the top-down information processing. The system has been implemented and evaluated on a humanoid robot. The experiments have shown promising results. Several interaction sessions were conducted with the robot. The robot was able to perceive the nonverbal cues of the interaction partner and behave accordingly.

Keywords: robotics, social robot, humanoid robot, perception system

1 Introduction

Analogous to humans, social robot should be able to produce and interpret verbal and nonverbal cues of the interaction partner. It is also essential to be able to show and recognize feedback during conversations. Feedback is an important aspect in maintaining the interpersonal communication and distributing the roles among people. Normally, people use feedback to select the suitable scenario in accordance with the situation and partner mood.

Numerous robot systems have been proposed so far that focus on the perception of the robot. The socially assisted robot PT1 from the project HOBBIT is developed as a care robot to observe humans in an indoor environment and interpret their actions [6]. Hobbit robot PT1 has a rich set of perception functionalities such as human detection, 3D human tracking, and action recognition. The robot was perceived mostly slow in the tasks, and the object learning process needs to be adjusted. However, the interaction with the robot is not intuitive, and it needs full consciousness for most of the gestures.

Nadine robot [9] is a highly realistic humanoid robot that exhibits certain social interaction capabilities. It can perceive human's body language for natural interaction using Microsoft Kinect RGB-D sensor. A gesture understanding human-robot interaction (GUHRI) system has been implemented in the robot that enables Nadine to understand and react to human gestures accurately and in real-time. The system is encumbered, which requires the human to wear physical assistive sensor. Although the system can perceive human gestures even with human-object interaction, it omits other interaction cues such as facial expression.

The humanoid robot ROMAN has been developed as a test platform for human-robot interaction [4]. The perception system of ROMAN has been initially developed by Schmitz [8], which involves auditory and visual perception. It is a sensor input driven hierarchy of perception modules, where each of these modules provides new information based on input images. Although the wide perception area in this robot, it lacks of interaction feedback perception. However, increasing the scope of psychological aspects involved in the concept design of these architectures can improve the human-robot interaction aspects.

This paper proposes a biologically inspired perception system of nonverbal feedback for human-robot interaction. Focusing on the vision modality, this perception system recognizes nonverbal feedback from the interaction partner during the conversation. The work is based on two psychological aspects: human perception and human's nonverbal feedback. The perception process in this work is based on the psychological theory of human perceptual cycle proposed by Neisser [5], whereas the feedback interpretation is relying on the psychological work of Allwood [3].

The rest of the paper is organized as follows: section 2 presents the proposed perception model of the robot. Section 3 shows the evaluation of the system. A conclusion is presented in the section 4.

2 The Proposed Perception Model

To develop a perception system with human-like capabilities, the psychological concepts and ideas in human perception should be considered. There are two main theories in the visual perception process of human: bottom-up and top-down information processing. Neisser [5] combines the two ideas to describe the process as a perceptual cycle. He believes that the perception is a continuous and cyclic process. The anticipatory schemata, the fundamental structure for vision, prepares the perceiver to accept certain kind of information. Therefore, the schemata direct the exploration process to look for the anticipated data. According to the exploration outcome, the original schemata is modified, and the process continues for more information. Figure 1 depicts the perceptual cycle.

The perception system described in this paper is based on the theory of perceptual cycle of Neisser [5]. It receives information from the sensor set to perceive and interpret the surrounding environment. The perception system consists of two main parts: schemata and exploration as shown in Figure 2.

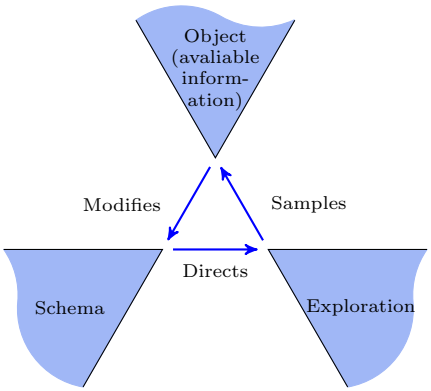


Fig. 1. The perceptual cycle [5].

2.1 Schemata

Schemata receives the information directly from the sensors and directs the perception process towards a particular type of information. It determines what will be perceived next according to the available information, memory, and previous knowledge. At each step, schemata constructs anticipations of particular kind of information that directs the exploration process. These anticipations occur at various levels of abstraction and meaning. Schemata consists of the following processes:

Low-Level Perception The low-level perception represents the extraction of low-level information for all modalities, referred to as *percepts*. These percepts are categorized to be interpreted in the next high-level step for a longer period. The low-level perception consists of several algorithms that have been used to detect nonverbal cues of the interaction partner. The face detection and head pose estimation of [7] has been used to recognize complex head movements. The facial expression recognition process of the proposed system is based on action units analysis. It uses deep learning as in [1]. To detect hand gestures, the proposed system used the work of [10].

High-Level Perception The received low-level percepts are accumulated over time to be interpreted within short periods. Head gestures and dynamic hand gestures are examples of these high-level percepts. To recognize dynamic head gesture, a sequence of head poses is compared with predefined sequences that represent some head gestures using Dynamic Time Warping (DTW) algorithm. The gesture with the minimum distance to the examined sequence is regarded as the winning gesture.

Tracking head poses of a human over a period enables of detecting head gestures such as nodding, shaking, tilting, and looking (gazing). Head gesture

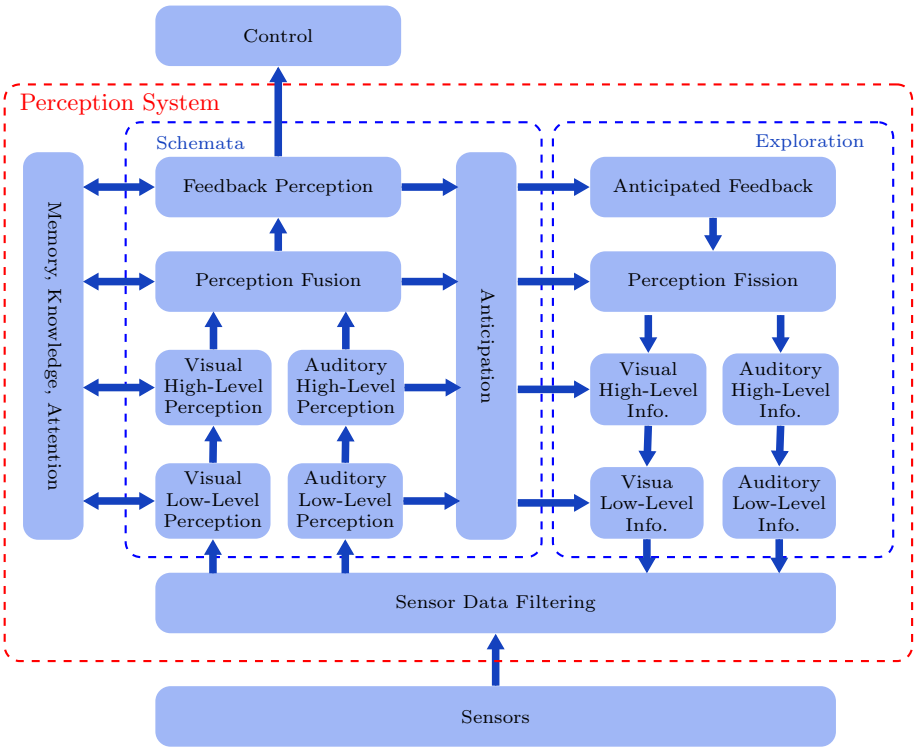


Fig. 2. Human perception model. A human perceives the surrounding environment in two phases: bottom-up and top-down information processing.

recognition has been implemented using the integrated Behavior-Based Control (iB2C) [2]. Implementing gestures as behaviors helps to select one gesture at a time and enables some gestures to inhibit other gestures which have different properties.

Perception Fusion All high-level percepts ordered chronologically are fused together to be interpreted later. The received percepts have different forms, and this stage categorizes them according to the subject belongs to rather than to their modalities. In this stage, percepts from some modalities may eliminate percepts from others.

Feedback Perception The fused percepts are then interpreted for a longer period. In this step, the given feedback by the interaction partner is interpreted according to the situation. The nonverbal cues are interpreted as feedback depending on a predefined tree that describes all feedback functions [2].

2.2 Exploration

Exploration process is directed by the schemata to focus on a specific kind of information during perception. Analogous to the schemata, exploration occurs on different levels of perception. Each level in exploration receives the anticipated information of the corresponding level. Exploration process achieves the following anticipations.

Anticipated Feedback The receiving part of feedback from the interaction partner activates prediction of the next nonverbal cues. This step determines the related cues that are required for specific feedback. These cues are transferred to the lower levels to get detailed information.

Perception Fission The required cues may rely on different modalities, which need to be split up into their corresponding sources.

High-Level Information This step filters the high-level information of each modality. According to this filtering process, the required low-level information will be sent to the lower level to focus on specific information.

Low-Level Information This step, according to the required information, specifies which information is needed to be acquired from the sensors. For example, this step can decide whether to focus on the facial expression of the interaction partner at the next moment or not.

3 Experimentation and Evaluation

In order to evaluate the proposed system, a gaming scenario is implemented in which human plays a game with a humanoid robot. In this section, we describe the game and the experimental platform and evaluate the performance of the system.

3.1 Experimental Setup

In order to validate our perception system, we have implemented a popular interactive game of 20 questions[4]. In this game, user has to think of something and then robot asks around 20 questions to guess that **thing**. Our version of game enables user to play it only through body gestures. Hand gestures are used for answering multiple choice questions. For **Yes** and **No** answers, head nodding and shaking can also be used. For every question, there are 5 possible answers: yes, no, unknown, irrelevant and sometimes.

[4] http://www.20q.net

Table 1. Confusion matrix of head gestures recognition. ND=Nodding, SH=Shaking, TL=Tilting left, TR=Tilting Right, LF=Looking Forward, LL=Looking Left, LR=Looking Right, LU=Looking Up, LD=Looking Down.

Detected as ⇒	ND(%)	SH(%)	TL(%)	TR(%)	LF(%)	LL(%)	LR(%)	LU(%)	LD(%)
Nodding	95.6	2.4	0.0	0.0	2.0	0.0	0.0	0.0	0.0
Shaking	2.3	90.2	2.2	2.0	3.3	0.0	0.0	0.0	0.0
Tilting left	0.0	0.0	95.8	0.0	2.9	1.3	0.0	0.0	0.0
Tilting right	0.0	0.0	0.0	93.0	4.4	0.0	2.6	0.0	0.0
Looking forward	0.0	0.0	1.9	1.1	92.7	0.0	0.0	4.3	0.0
Looking left	0.0	4.1	0.0	0.0	0.0	92.1	0.0	3.8	0.0
Looking right	0.0	2.9	0.0	0.0	0.0	0.0	94.7	2.4	0.0
Looking up	1.8	0.7	0.0	2.3	0.0	0.0	0.0	95.2	0.0
Looking down	2.5	0.0	0.0	0.0	3.6	1.2	0.0	0.0	92.7

The proposed system has been implemented in the humanoid robot ROBIN of the University of Kaiserslautern. ROBIN is equipped with a backlit projected face, arms, hands and torso. The face makes use of projective technology to express almost any facial expression using action units. ASUS Xtion Pro is installed on the chest of robot. The whole arm has 14 degrees of freedom, where hands are able to perform nearly all gestures. For perception, a standalone system, Intel Core i7 running at 3.40 GHz, has been attached to ROBIN to process the RGB-D data.

3.2 High-Level Perception Evaluation

Before evaluating the whole perception system, the high-level perception could also be evaluated. This experiment presents the recognition and rough interpretation of high-level cues. This experiment is conducted individually and not in a context of interaction. Therefore, there is no need to interpret these gestures on the psychological basis.

Different head gestures were recorded for 16 participants. The participants have been told to express different gestures in front of the robot. The experiment shows that the high-level perception process able to detect head gestures accurately, which can boost the perception of human feedback. Table 1 shows the confusion matrix of the recognition rate of head gestures, which is in overall more than 90%. The confusion matrix shows the overlap among different gestures. The overlap between static and dynamic gestures usually occurs when the dynamic gesture is too slow or too fast. Slow dynamic gestures are more recognizable than fast ones, and static gestures are more stable than dynamic gestures.

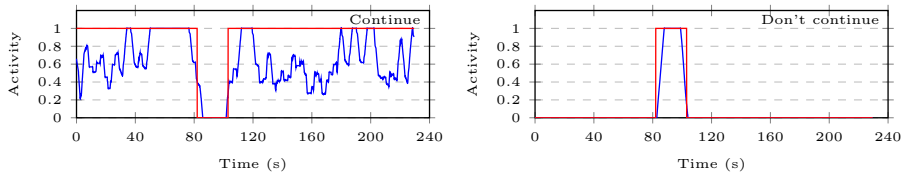
3.3 Feedback Perception Evaluation

For evaluation, five different subjects play 20 questions game with the robot. As the subjects don't know how to play the game, robot explains the game at the beginning. To answer multiple choice question, basic hand gestures are used.

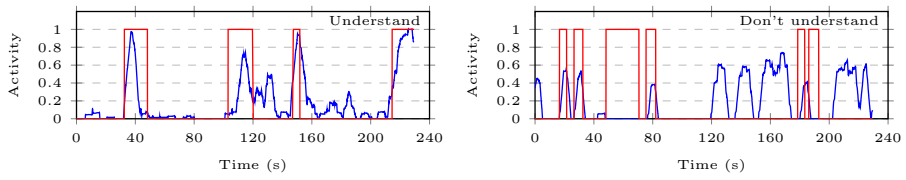
Human information related to the feedback, such as the activation of all the behaviors, face expressions, head, and hand gestures along with the question are stored for every frame to be used later in the analysis. The playing sessions are also recorded to be analyzed by an expert. He observes the feedback behaviors whenever a question has been asked by the robot and fills in a questionnaire. He also takes notice of subtle changes in the behavior, for example, lack of interest, understanding or not understanding behaviors, etc. At the end of the experiments, the recorded information by the system is compared with the expert questionnaire. Figure 3 shows plots of different behaviors recognized by the system (blue line) as well as behaviors detected by a human expert (red line).

The system has demonstrated the recognition of key moments (activations) in all behaviors, which can be seen from the plots. In Figure 3a, which shows whether the partner is interested in the game, the system recognized that the behavior *don't continue* is active between time 80*s* and 100*s*. Exactly in the same time frame, the human expert also detected the activity of this behavior. Furthermore, the system follows the human results regarding the behavior *continue*. An important statistic can be seen in this figure that even a small value of activeness of this behavior, in this case 0.2, suggests that this behavior is active according to the human expert observation.

However, in Figure 3b, the activity of *understand* behavior above 0.6 is recognized as understanding signal. While system missed *don't understand* behavior couple of times and activate this behavior at wrong time frames. Since humans can detect small subtle changes with ease, hence faint changes in the appearance, in this case eyebrows, are detected in contrast to our facial expression module.



(a) Continuation



(b) Understanding

Fig. 3. Feedback types. Comparison between system activity and human observations for two feedback types. Blue line represents the system values and red line represents the expert observations.

4 Conclusion

Human perception system has been studied by psychologists extensively, and many theories have been proposed. The present work is a perception system for a social robot that can interact with humans. The proposed system is based on the perceptual cycle theory presented by Neisser. The system is composed of two phases: bottom-up and top-down information processing. In the bottom-up phase, nonverbal feedback is derived starting with sensory information passing different levels of data abstraction. It comprises of low-level, high-level, and feedback perception. Contrary to the bottom-up perception, top-down perception uses the contextual information to guide the perception process. Starting from the received nonverbal cues and the expected feedback, this stage focuses on a specific kind of information that is relevant to the ongoing event. Propagation of the required information from the higher level to the lower level enables seeking specific information. Omitting irrelevant information in each step reduces the information space tremendously and leads to less processing time.

References

1. Al-Darraji, S., Berns, K., Rodić, A.: "Action Units Based Facial Expression Recognition Using Deep Learning", in: Advances in Robot Design and Intelligent Control: Proceedings of the 25th International Conference on Robotics in Alpe-Adria-Danube Region (RAAD), Belgrade, Serbia (2016).
2. Al-Darraji, S., Zafar, Z., Berns, K.: "Real-time Perception of Nonverbal Human Feedback in a Gaming Scenario", in *Proceedings of the 2016 British HCI Conference*. Bournemouth, UK, July 11–15 2016.
3. Allwood, J.: "Cooperation and Flexibility in Multimodal Communication", in *Cooperative Multimodal Communication*, ser. Lecture Notes in Computer Science, H. Bunt, R. Beun, Eds., Springer Berlin Heidelberg, 2001, vol. 2155, pp. 113–124.
4. Berns, K., Schmitz, N.: "Perception System for Naturally Interacting Humanoid Robots", *Künstliche Intelligenz (KI)*, vol. 1/09, 2009.
5. Neisser, U., *"Cognition and reality : principles and implications of cognitive psychology"*, W.H. Freeman San Francisco, 1976.
6. Papoutsakis, K., Padeleris, P., Ntelidakis, A., Stefanou, S., Zabulis, X., Kosmopoulos, D., Argyros, A.: "Developing visual competencies for socially assistive robots: the HOBBIT approach", in *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, ACM. 2013, p.56.
7. Saleh, S., Berns, K.: "Nonverbal Communication With a Humanoid Robot Via Head Gestures", in *PETRA '15: Proceedings of the 8th International Conference on PErvasive Technologies Related to Assistive Environments*. Corfu, Greece: ACM, USA, July 1–3 2015.
8. Schmitz, N.: "Dynamic Modeling of Communication Partners for Socially Interactive Humanoid Robots", Dissertation, University of Kaiserslautern, 2011.
9. Xiao, Y., Liang, H., Yuan, J., Thalmann, D.: "Body Movement Analysis and Recognition", in *Context Aware Human-Robot and Human-Agent Interaction*, Springer, 2016.
10. Zafar, Z., Berns, K.: "Recognizing Hand Gestures for Human-Robot Interaction", in *Proceedings of the 9th International Conference on Advances in Computer-Human Interactions (ACHI)*. Venice, Italy, 2016.