Gaze and Pointing: Two Basic Nonverbal Expressions Enable Embodied AI to Communicate

Feiyang Xie
Yuanpei College
Peking University
2100017837@stu.pku.edu.cn

Abstract

Communication is one of the important abilities that makes humans far superior to other animals. Although verbal communication tends to be primary in human-human interactions, nonverbal behaviors, such as eye gaze and gestures, can convey mental state, augment verbal communication, and reinforce what is being said. Moreover, relevant research shows that nonverbal communication may promote language production, thereby promoting communication development. As embodied robots become more humanlike in appearance and more prevalent in society, it is crucial that they master some nonverbal expressions to improve communication efficiency and accuracy between humans and robots and reach consensus in specific scenarios. Among the nonverbal behaviors used daily, gaze and pointing are two of the most frequently used and efficient for communication. We can infer gaze and pointing are also important in human-robot interaction (HRI). This essay will first briefly introduce gaze and pointing and discuss their indispensability in human-robot communication. After this, we will analyze how to learn and use gaze and pointing to achieve efficient communication in an indoor, one-on-one human-robot interaction setting. Finally, we will discuss the important nonverbal expressions robots need to achieve more universal human-robot communication as embodied robots continue to develop.

1 Introduction

Communication is an important way for humans to share and obtain information, and it is also one of the important abilities that make humans far superior to other animals. In most cases, communication can be divided into two types: verbal and nonverbal, and nonverbal communication usually carries most of the information. In addition, nonverbal expressions are often more versatile than verbal expressions. For example, when a person travels to a place where they do not speak the language, they can use gestures resembling the object they refer to in order to convey their message to others. Among all nonverbal expressions, gaze and pointing are used most frequently in daily life because they are easy to implement and efficiently convey information [1]. In the possible future society where humans and robots coexist, it is crucial for efficient communication and consensus between humans and robots. At the level of language communication, there has been much research and great progress, such as language models like GPT-4 and vision-language models like GPT-4V [2]. Due to the abstract and context-related nature of nonverbal communication, current robots cannot learn and use human nonverbal expressions well. To build a more versatile robot that meets human needs, the robot must have nonverbal expression abilities. Starting with the most basic and practical nonverbal expressions is very meaningful, as it not only reduces problem difficulty but also enables more complex subsequent research. Gaze and pointing are ideal candidates. In the following sections, I will introduce gaze and pointing and analyze their role in human-robot communication. In Sec. 4, I will discuss learning and using gaze and pointing during communication between a robot and human in a specific environment. Finally, we will discuss the important nonverbal expressions robots need to achieve more universal human-robot communication as embodied robots continue to develop.

2 Gaze

Gaze is important to human-human interactions because it is closely tied to what people are thinking and doing. During communication, eye gaze can convey information, regulate social intimacy, and express social or emotional states. Moreover, people use observations of others' eye gaze to guide communication [3]. There is much evidence for the specificity and effectiveness of gaze. Generally, gaze is directed at conversational partners approximately 60-80% of the time [4]. Biological research shows the human eye structure differs greatly from other animals. The white sclera occupies most of the human eye area, highlighting the pupil and making its direction easier to perceive [3]. This evolution enables the social communication of gaze, as humans readily recognize eye gaze and infer intentions from it. In a specific environment, gaze can represent diverse information and has many types. For example, gazing at a certain place demonstrates a person's focus, while eye movements may indicate desired items. According to [5], gaze is categorized as: Mutual gaze, Referential/Deictic gaze, Joint attention, and Gaze aversions.

Gaze plays a huge role in communication between people, but in the process of communication between people and robots, these functions are not all required. I think that when people communicate with embodied robots, the gaze learned and used by robots should include the following functions:

1. **Object reference** Object reference is the most basic role that gaze plays in communication. In indoor environments, when people mention items that exist in the scene, they tend to look in the direction of the items, especially items that they think the person they are communicating with has not seen. Therefore, during communication, the robot needs to use gaze to refer to some items that are highly relevant to the communication content. After this, the robot tries to have a mutual gaze with human to confirm that they agree on the object they are referring to.
2. **Attention detection** In communication, attention determines the environmental information and contextual information obtained by the communicator, and humans can often determine the scope of attention through gaze. Therefore, robots need to have the ability to capture human attention by their gaze, so that communication can be better carried out in the human-focused way. In addition, robots also need to learn how to change human attention.
3. **Prompt next action** Gaze between humans is dynamic and follows the task at hand. For example, the "reference-action sequence"—in which an instructor refers to an object and then a worker acts on that object-can be divided into five cyclically repeating phases, each with their own distinct gaze behaviors: pre reference, reference, post-reference, action, and post-action. Therefore, when a robot assists a human to complete a certain task, it needs to understand the human's next action intention and convey its own next action intention through gaze [6].

3 Pointing

Pointing is another important universal nonverbal expression in humans. In every studied human culture, infants typically begin pointing between 9-14 months, arguably producing their first purely informative gesture. Related research shows children delayed in pointing are also delayed in subsequent language acquisition. This demonstrates pointing strongly correlates with language and plays a vital communication role. There are many hypotheses about the origin of pointing, with studies indicating it likely extended from touch [7].

In daily life, the meaning of pointing is highly related to environmental and contextual information, and is often used together with gaze. Although pointing is very similar to gaze, gaze cannot replace pointing, and embodied robots should learn and use pointing. The following are the main aspects that pointing surpasses gaze:

1. **More precise reference** In a complex environment, humans tend to use pointing to convey reference to objects, because compared to gaze, pointing has a larger range of movements, is easier to perceive, and has higher direction and accuracy. Therefore, the robot needs to be able to evaluate the current communication and environment to decide which nonverbal expression to use to refer to the object.

2. **Describe non-salient objects at hard-to-describe positions** In communication, pointing can often be used to describe things that are difficult to describe through gaze. Therefore, the information that pointing can carry is often richer and more abstract. For example, we can refer to the sky or stars by pointing to the ceiling, the weather or scenery outdoors by pointing to the windows.
3. **Show strong desire and intention** Based on the assumption that pointing is an extension of touch, we can say that pointing is usually accompanied by strong desire and intention. In communication, gaze usually can only express the shift of attention, and the ambiguity is high in conveying whether the object being looked at is needed, while pointing can reduce the ambiguity of intention.

4 How to learn and use gaze and pointing?

4.1 Learn gaze and pointing

The abstract nature and diversity of nonverbal expressions make them very difficult for robots to correctly learn and use. There has been much research on robots utilizing gaze, largely divided into three categories: human-focused, design-focused, and technology-focused. These studies have achieved good results in certain aspects. For instance, human perception studies have found people can successfully identify a robot's gaze target, whether looking at them or other objects. Design-focused studies show robot gaze can improve human-robot interactions across domains, with mutual gaze and gaze aversions regulating conversational pace and participation [5]. However, in these studies robots lacked sufficient understanding of human gaze. More importantly, properly using nonverbal expressions requires cognitive common sense. Some studies have tried modeling mental states behind expressions, like understanding pointing by constructing a relevance-based theory of mind model [8]. Results indicate pointing's effectiveness stems from theory of mind inferences supporting relevance calculation, demonstrating learning pointing can truly improve communication. However, for embodied robots to use gaze and pointing more appropriately, they need to learn from demonstrations via imitation and contrastive learning. This can better teach correlations between expressions and context.

4.2 Use gaze and pointing in communication

In my opinion, nonverbal expressions' main functions are (1) reducing communication cost and improving efficiency, and (2) conveying information difficult to describe verbally. Therefore, robots should use nonverbal expressions or combine them with verbal ones based on those two situations. Based on the above analysis, we need a cost measurement model, with Rational Speech Act (RSA) a strong candidate [9]. To simplify, we will examine a one-on-one indoor human-robot communication setting. For these two nonverbal expressions, their use can be roughly divided into two time frames: during communication and at specific moments. For instance, gazing at the speaker improves communication efficiency at minimal cost and occurs during the interaction. Pointing towards a direction to focus a human's attention corresponds to the second case. The whole communication process can be abstracted as: for a given context, the robot infers its intentions and generates all possible expressions, scoring them on cost, efficiency, accuracy and other metrics. The top-scoring expression is the final choice.

5 Conclusion

Although gaze and pointing play important communication roles, using just these two nonverbal expressions cannot achieve human-level efficiency. Studies show observers often misinterpret pointers' indicated locations [10]. For more accurate and efficient communication, we need to introduce more complex nonverbal expressions like gestures and symbols. However, for robots to understand and correctly use such complex nonverbals, further advances are required in machine common sense and cognitive modeling. There remains a long path before robots can communicate like humans.

References

[1] Sotaro Kita. *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003. 1

[2] OpenAI. Gpt-4 technical report, 2023. 1

[3] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000. 2

[4] Michael Argyle and Roger Ingham. Gaze, mutual gaze, and proximity. 1972. 2

[5] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017. 2, 3

[6] Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3603–3612, 2015. 2

[7] Cathal O'madagain, Gregor Kachel, and Brent Strickland. The origin of pointing: Evidence for the touch hypothesis. *Science Advances*, 5(7):eaav2558, 2019. 2

[8] Kaiwen Jiang, Stephanie Stacy, Annya L Dahmani, Boxuan Jiang, Federico Rossano, Yixin Zhu, and Tao Gao. What is the point? a theory of mind model of relevance. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022. 3

[9] Judith Degen. The rational speech act framework. *Annual Review of Linguistics*, 9:519–540, 2023. 3

[10] Oliver Herbort, Lisa-Marie Krause, and Wilfried Kunde. Perspective determines the production and interpretation of pointing gestures. *Psychonomic Bulletin & Review*, 28:641–648, 2021. 3