Incorporating Nonverbal Expressions into An Embodied Chat Robot

Xinyi Yang
Department of Automation
Tsinghua Universtiy
xy-yang21@mails.tsinghua.edu.cn

Abstract

Nonverbal expressions which can convey rich social information enable humans more suitable and effective communication. Not only are these expressions deeply rooted in human cognition, but they are also important in Human-Robot Interaction for enhancing trust and familiarity. There are already researches about several nonverbal expressions like gaze and point. However, most existing works focus on one single expression or one specific application. Thus we propose an architecture based on "five minds" to model the process of communication between two agents, which may serve as an inspiration for building an embodied chat robot as our wish.

1 Introduction: Visual Modes of Communication

People use a wide range of communicative acts across different modalities, from tiny facial expressions to exaggerated body movements, or from direct demonstrations to abstract language. Among various forms of communication, nonverbal visual modes are the most ubiquitous in daily life. From ancient etchings on cave walls to modern digital displays, the ability to externalize our thoughts in visual form lies at the heart of key human innovations and forms the foundation for the cultural transmission of knowledge.[7] More naturally, whether it is a gaze or a pointing action, we can express intentionally or unintentionally, and capture what the other person wants to express through these visual changes. For example, four-month-old infants are known to use eye gaze cuing to help visually process objects (*e.g.*, Striano and Reid [21]); At as young as the age of one, infants use pointing to inform another person of the location of an object that person is searching for (*e.g.*, Liszkowski et al. [16]). On top of that, when in a foreign country where you don't understand the local language, visual modes become the universal language.

There's no doubt that nonverbal visual language is essential for human society, but it also plays an important role in Human-Robot Interaction (HRI) especially for embodied agents. On the one hand, there is always ambiguity in human words, which can be interpreted more clearly with the help of visual language. For example, when a person gives an instruction "Can you please get this for me?" to a robot with pointing to a pen, the robot will easily understand what "this" refers to without asking the person. And when a person displays an impatient expression but has said nothing, it will be better for the robot to capture the impatience and leave her or him alone. On the other hand, incorporating responsive, meaningful and convincing nonverbal expressions into robotics helps the humanoid agent to engender the desired familiarity and trust, and makes HRI more intuitive.[9] Several modern-day approaches are trying to manipulate features of robot appearance and behavior and measure their influence on human responses (*e.g.*, Ito et al. [11] found that when a robot is learning from human demonstration, displaying mutual gaze leads people to view the robot as more intentional than displaying random gaze; people spend more time teaching the robot pay more attention to it, and speak more with it).

How to model such nonverbal expressions has been under long-time research. And the combination of verbal and nonverbal communication is also one of the central topics in HRI. However most existing work focuses more on one single kind of nonverbal language such as gaze and point, or applies the combination to acquiring specific information such as reference[5]. To answer the question of how to design a general embodied chat robot, this essay is organized as: Firstly, as primitive and crucial parts of human communication, gaze and point will be reviewed respectively about the roles in Human-Human Interaction (HHI) and computational models in Sec. 2, to glance at studies of non-verbal expressions. Then in Sec. 3, an architecture of the process that two agents communicate with both verbal and nonverbal language will be proposed. Finally, a discussion about current limitations and future directions will be made in Sec. 4.

2 Literature Review: Non-verbal Language in both Humans and Robots

According to Kendra Cherry [14], research has identified 8 types of nonverbal communication:

- Facial expressions
- Gestures
- Paralinguistics (such as loudness or tone voice)
- Body language
- Proxemics or personal space
- Eye gaze
- haptics(touch)
- Appearance
- Artifacts (objects and images)

Examples of utilizing these types of nonverbal communication to convey and get underlying attention, intention, emotion and personality can be found everywhere in everyday life. People are often good at identifying what expressions through nonverbal language mean. To model the process of common human communicating, it is necessary for us to figure out how nonverbal language works in HHI and what current computational models can do about nonverbal interactions in HRI. Thus, *gaze* and *point* will be present below for their uniqueness that distinguishes humans from other animals and their importance in human communication.

2.1 Gaze

2.1.1 Gaze Communication in HHI

Evidence from psychology suggests that eyes are a cognitively special stimulus, with unique "hard-wired" pathways in the brain dedicated to their interpretation.[6] And given the definition in Emery [6], social gaze is not limited to information from the eyes. The whole head, in particular the orientation in which it is directed (using the nose, for example) is a sufficient indicator of attention direction. For one thing, gaze communication allows people to communicate with one another at the most basic level regardless of their familiarity with the prevailing verbal language system. For another during conversations, gaze can be used to convey information, regulate social intimacy, manage turn-taking, control conversational pace, and convey social or emotional states.[15] Moreover, the information behind the gaze can also be used to predict what the other people is going to say.[4]

According to Fan et al. [9], atomic-level gaze communications are distinguished into six classes: (1) *Single* refers to individual gaze behavior without any social communication intention; (2) *Mutual* occurs when two agents look into eyes of each other to establish a communicative link; (3) *Avert* happens when gaze of one agent is shifted away from another to express distrust, fear or thoughtfulness; (4) *Refer* tries to induce another agent’s attention to a target; (5) *Follow* happens when one agent perceives gaze from another and follows to contact with the stimuli the other is attending to; (6) *Share* means appears when two agents are gazing at the same stimuli. It is long term, coarse-grained temporal composition of the above atomic-level patterns that usually appears at the event-level in real situations.

2.1.2 Gaze Communication in HRI

Based on these phenomena displayed in human gaze communication, modern-day approaches to incorporating gaze into HRI vary widely. Combining the idea of Admoni and Scassellati [1] and our focus on building an embodied chat robot, we divide the corpus of work on gaze in HRI into three categories both by their goals and methods. These categories are as follows:

Human-focused aims to perceive and understand gaze during HHI and HRI. Given head location, Recasens et al. [18] proposed a deep learning based model for extracting head pose and gaze orientation, following the gaze of the person and identifies the object being looked at in the image. Then they further extended their work to videos.[19] These works only focus on predicting single-person gaze. Then Fan et al. [8] proposed a spatial-temporal neural network to detect shared attention intervals in videos and predict shared attention locations in frames. They also extended their work to understanding gaze communication in social videos from both atomic-level and event-level.[9]

Design-focused investigates how design choices about a robot, such as its appearance or behavior, can impact interactions with humans.

Technology-focused aims to build computational tools for generating robot eye gaze in HRI. Current approaches extend from biologically-based models (*e.g.*, ACT-R/E in Trafton et al. [24] performs conversational tracking by switching its visual attention to the speaker in a multi-party conversation) to empirical models (*e.g.*, Admoni et al. [2] extracted statistical information on timings and directions of gazes in dyadic conversations that achieve certain conversational functions) to heuristic systems (*e.g.*, a parametric computational model for animating gaze shifts of virtual agents is successful at performing gaze shifts to peripheral targets[3]).

2.2 Point

2.2.1 Point Communication in HHI

All animals including humans understand the world by collecting observations through individual perception. Being social creatures, however, we also get observations that are pointed out to us by others. The key characteristic of pointing is that it leads to joint attention.[13] When a signaler points an object to a receiver, she invites the receiver to become a "guest" to the observation. The receiver can safely assume that the information given by the "host" must be relevant to the shared task. Thus a crucial function of pointing is always paternalistic helping, which has been found prevalent in children.[12] It is also noted that pointing is *overloaded* (*i.e.*, the same pointing gesture has many interpretations, making the referent of pointing ambiguous when considered in isolation) and *indirect*. [12]

2.2.2 Point Communication in HRI

The combination of verbal and nonverbal communication for reference is one of the central topics in HRI, among which the most obvious and straightforward nonverbal way is pointing. There are researches that emphasize pointing direction and thus are not object-centric while missing language reference (*e.g.*, Shukla et al. [20] presented a probabilistic and appearance-based object detection framework to detect pointing gestures and robustly estimate the pointing direction, and proposed a functional model for both finger pointing and tool pointing using an object in hand). Furthermore, there are also researches that focus on the understanding of embodied reference (*e.g.*, YouRefIt, a new crowd-sourced dataset of embodied reference collected in various physical scenes, is introduced in Chen et al. [5]. They further devise two benchmarks for image-based and video-based embodied reference understanding).

3 Architecture: Communication between Two Embodied Agents

After a brief review about two typical nonverbal expressions, gaze and point, what should be considered next is how and when will it combine or alternate verbal and nonverbal expressions. Based on the natural observation of human embodied communication, we separate the process into two stages, thinking stage (*i.e.*, update mental states with information acquired from own observation and other people's expressions) and expressing stage (*i.e.*, choose and conduct appropriate expressions according to current mental states). In order to simplify, these two stages proceed alternately. To elaborate the proposed two-stage architecture, the process of communication between two embodied agents will be taken as an example to illustrate how these stages work.

3.1 Thinking Stage

3.1.1 Mind Representation

To account for modeling the process of building and updating mental states during the communication between two embodied agents, we adopt a novel structural mind representation termed "five minds" proposed by Fan et al. [10]. The representation includes two first-order self mental states (*i.e.*, the ground-truth mental state, representing own cognition towards the world), two second-order estimated mental states (*i.e.*, an estimation of others' ground-truth mental state, which may deviate from the ground-truth mental states), and the third-level "common mind". The proposed "five minds" does not attempt to infer mental states among agents recursively with potentially infinite loops, instead the "common mind" considers what the two agents share completely transparently without infinite recursion and corresponds to the concept of "common ground"[23].

3.1.2 Update Rules

Based on "five minds", the question becomes updating every mind state according to the environment and information from both verbal and nonverbal expressions. Apparently every agent's own observation of the environment changes own first-order self mental state, and clear information conveyed through verbal expressions changes the other agent's second-order estimated mental state and thus may cause changes to the "common mind". As for nonverbal expressions, they may serve as clues of the ambiguity in verbal expressions, such as "this/that", "she/he" and other pronouns, which may be the reference of a pointing gesture. They may also serve as direct information for updating each other's second-order estimated mental state and thus the "common mind". For instance, gaze communication uses eye gazes as portals inward to provide agents with glimpses into the inner mental world.[9] Moreover, nonverbal expressions may even affect first-order self mental states through guided observation. As for gaze, an agent may get more observations after following the other agent's gaze. (Fan et al. [10], Fig. 1) Similar evidence can also be found in "joint attention" mentioned above in Sec. 2.2.

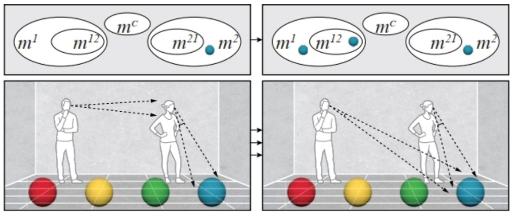

Figure 1: "five minds" in following attention.

3.2 Expressing Stage

What expressions to choose for better communication is based on current mental states. There are two main factors to consider. The first is for more suitable communication, *i.e.*, choosing what the environment allows and the other agent prefers. According to own first-order mental state, for example, if everyone else around is quiet then it is more suitable to choose nonverbal expressions, or if the other agent isn't paying attention to your body then verbal expressions are more suitable. Furthermore, according to second-order estimated mental states, the other agent's preferred expressions of a specific object or event can be captured. Using these expressions will make the other agent feel more comfortable.

The second is for more effective communication, which is of great importance in goal-driven situations. Under most circumstances, what sparks communication is one agent trying to teach the other agent something, *e.g.*, the location of an object (concrete), a mathematical concept (abstract), or whatever. Sumers et al. [22] reported a pair of studies that compare *demonstration* (a form of example-based teaching, *e.g.*, showing someone how chess pieces move) and *language* (convey information about categories, relations and causal structures), which both are fundamental means of knowledge transmission. Results show that language outperforms demonstration when communicating complex concepts, because language relies on shared abstractions to efficiently transmit complex concepts and demonstrations struggle to communicate such concepts but are less reliant on shared abstractions.

4 Discussion: More Nonverbal Expressions and Their Integration

There are still limitations in current researches to implement the architecture proposed above. From my standpoint, this architecture calls for modeling more expressions except gaze and point and integrating these expressions in the embodied environment. For example, facial expressions, volume, and tone, are critical for updating second-order estimated mental states. The reason why we feel GPT[17] is not very clever is there is only language modality included. We hope that this architecture will inspire future researches to integrate these modes into an embodied chat robot.

References

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

[2] Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati. Are you looking at me? perception of robot attention is mediated by gaze type and group size. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 389–395. IEEE, 2013. 3

[3] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. A head-eye coordination model for animating gaze shifts of virtual characters. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, pages 1–6, 2012. 3

[4] Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerard Bailly, Frederic Elisei, Sascha Fagel, Peter Ford Dominey, and Jocelyne Ventre-Dominey. I reach faster when i see you look: gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in neurorobotics*, 6:3, 2012. 2

[5] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1385–1395, 2021. 2, 3

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

[7] Judith E Fan, Robert D Hawkins, Mike Wu, and Noah D Goodman. Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, 3:86–101, 2020. 1

[8] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018. 3

[9] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019. 1, 2, 3, 4

[10] Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. Learning triadic belief dynamics in nonverbal communication from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, 2021. 4

[11] Akira Ito, Shunsuke Hayakawa, and Tazunori Terada. Why robots need body for mind communication-an attempt of eye-contact between human and robot. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*, pages 473–478. IEEE, 2004. 1

[12] Kaiwen Jiang, Stephanie Stacy, Chuyu Wei, Adelpha Chan, Federico Rossano, Yixin Zhu, and Tao Gao. Individual vs. joint perception: a pragmatic model of pointing as communicative smithian helping. *arXiv preprint arXiv:2106.02003*, 2021. 3

[13] 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

[14] MSEd Kendra Cherry. Types of nonverbal communication. `https://www.verywellmind.com/types-of-nonverbal-communication-2795397`. 2023.11.4.

[15] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

[16] Ulf Liszkowski, Malinda Carpenter, Tricia Striano, and Michael Tomasello. 12-and 18-month-olds point to provide information for others. *Journal of cognition and development*, 7(2): 173–187, 2006.

[17] OpenAI. Gpt-4 technical report, 2023.

[18] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015.

[19] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017.

[20] Dadhichi Shukla, Ozgur Erkent, and Justus Piater. Probabilistic detection of pointing directions for human-robot interaction. In *2015 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–8. IEEE, 2015.

[21] Tricia Striano and Vincent M Reid. Social cognition in the first year. *Trends in cognitive sciences*, 10(10):471–476, 2006.

[22] Theodore R Sumers, Mark K Ho, Robert D Hawkins, and Thomas L Griffiths. Show or tell? exploring when (and why) teaching with language outperforms demonstration. *Cognition*, 232: 105326, 2023.

[23] Michael Tomasello. *Origins of human communication*. MIT press, 2010.

[24] J Gregory Trafton, Magda D Bugajska, Benjamin R Fransen, and Raj M Ratwani. Integrating vision and audition within a cognitive architecture to track conversations. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 201–208, 2008.