Methods and Challenges for Enabling Embodied AI to Communicate with Gestures

Ningxin Pan
Yuanpei College
Peking University
2100017816@stu.edu.cn

Abstract

Non-verbal communication plays a significant role in human interactions, and whether agents can utilize this form of communication is an important question. This article focuses specifically on the gesture aspect of non-verbal communication and analyzes it from the following perspectives 1) Several aspects to consider when building an agent able to communicate with gesture (Section 2) ; 2) Several challenges when building this kind of agent (Section 3).

1 Introduction

Artificial Intelligence (AI) has become a cornerstone of modern technology, permeating nearly every aspect of our lives. Its influence extends beyond the realms of data analysis and automation, seeping into the intricate domain of human behavior, particularly nonverbal communication. The role of AI in decoding nonverbal communication is an emerging field of study that holds promising potential for a multitude of applications.

Nonverbal communication, which includes facial expressions, body language, and tone of voice, is a critical component of human interaction. It often conveys more information than verbal communication and is instrumental in understanding the nuances of human behavior.

This article will focus primarily on gesture analysis, as making facial expressions and changing intonation is still too challenging for agents.

2 How to Enable Agent to Communicate with gestures

2.1 Imitation

Manually designed gestures can be inspired by observing recorded human gestures and recreating them using key frames that suit the robot's physical properties. [4] (Fig. 1) Another approach is to record human motion and automatically translate it onto the robot, accounting for differences in morphology. This process, known as learning from demonstration or imitation learning, has been used to teach robots various tasks and for telepresence applications.

Traditionally, motion capture technologies with complex multi-camera setups were used to record human motion. However, more compact depth sensors, like Microsoft's Kinect, allow for markerless tracking with a single portable camera, albeit with some loss of recording accuracy. Recently, advanced AI and computer vision techniques have enabled the extraction of three-dimensional motion recordings from data found "in the wild," such as YouTube videos. [7]

Once recordings of human-performed gestures are collected, mapping or translation of these recordings is necessary due to differences in size and kinematic abilities between the original performer and the robot. [3] Algorithmic calculations or neural network-based approaches are used for this mapping process.

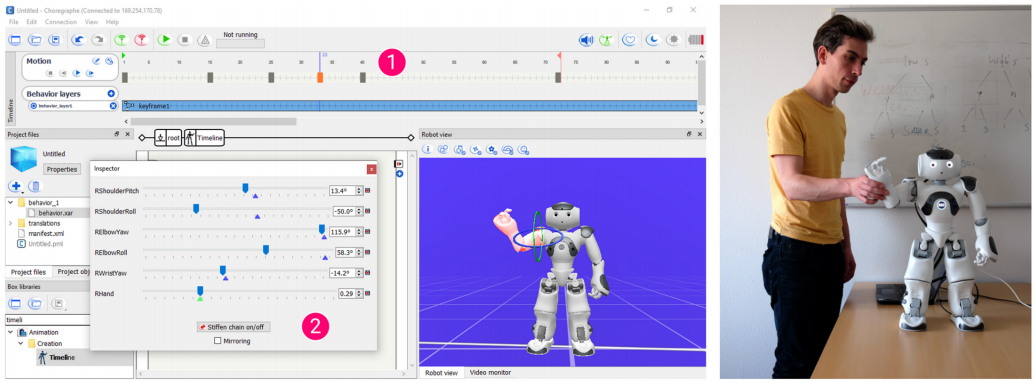

Figure 1: Manual gesture design by defining key frames

Learning from demonstration is less labor-intensive compared to manual gesture design since once the mapping is established, multiple recordings can be transferred to the robot. Variations and detailed frame-by-frame definition can be easily introduced, and aspects like intention and enthusiasm observed in human gesturing behavior can be incorporated directly into the robot's performance. However, it is challenging to account for differences in physical appearance and kinematic abilities between humans and robots, resulting in a loss of detail when translating gestures. Imperfections in capturing or mapping can also introduce unintended movement or jerkiness. Applying post-processing steps, such as denoising and dimensionality reduction [8] , can help create smoother gestures.

2.2 Combine Verbel Information

After determining which gesture to perform, the motions also need to be timed in such a way that they correspond to the related information that is conveyed through the robot's verbal information.

Once the appropriate gesture has been determined, it is essential to time the motions in a way that corresponds to the related information conveyed through the robot's speech. As it is shown in Fig. 2, this can be achieved through trial-and-error or manual annotation when the robot's output is known beforehand. [1] Timing can also be based on heuristics derived from human gesture studies, although these may not consider differences between humans and robots in terms of execution speed.

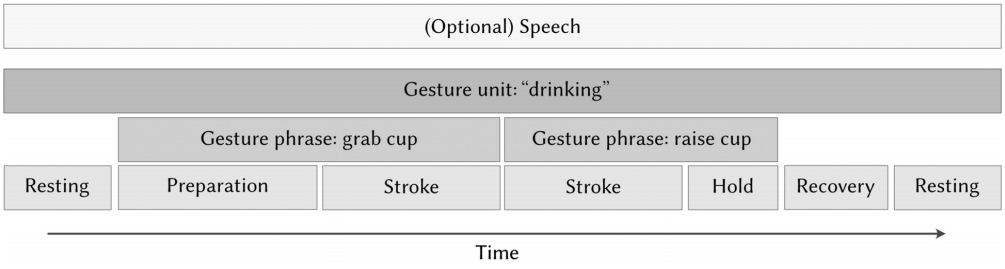

Figure 2: Breaking down a gesture unit into its constituent parts: different phrases and multiple phases

To account for the robot's physical abilities, some implementations prerender the robot's utterance to an audio file via text-to-speech before playing it. [6] This allows for more accurate alignment of co-speech gestures by combining knowledge of the robot's motor speeds with the timing required for pronouncing certain words. [2]

Contextual factors, such as aligning the gesture stroke with the most accentuated syllable in a phrase or handling overlapping or interrupted speech, can also influence the optimal timing of gestures. In the field of multi-modal output generation for virtual agents, various behavior realization methods have been proposed to facilitate gesture production and co-speech timing. These methods involve stretching or shrinking gestures, connecting individual gestures to create a smooth overall motion, and utilizing animation techniques such as motion capture, key frame animation, and physical simulation. [9]

3 Challenges for Agent to Communicate with Gestures

Developing AI systems that can communicate with human gestures poses several challenges. I roughly categorize them into two groups.

3.1 Challenges Related to the Traits of Human Gestures

As for the fact that human gestures are always based on some common sense [5], they can vary greatly based on cultural, individual, and contextual factors. Explaining and understanding this variability requires artificial intelligence systems to recognize and adapt to a wide range of gestures. Additionally, gestures can be complex, involving subtle nuances and combinations of movements, making it challenging for AI systems to accurately interpret their meanings.

Gestures often derive meaning from the context in which they are used. [10] For example, the same hand movement can have different interpretations in different situations. AI systems need to have contextual awareness to correctly understand and interpret gestures in a given environment while considering accompanying speech, environmental cues, and social dynamics.

Gestures are often used in conjunction with other modes of communication, such as speech, facial expressions, and body language. Deciphering the intent behind a gesture may require analyzing multiple modalities simultaneously. Combining these different streams of information and resolving ambiguity poses a complex task for AI systems.

3.2 Challenges Related to the System processing

Building effective AI models for gesture recognition and interpretation requires diverse datasets capturing the variability of gestures. However, due to the subjective nature of gesture interpretation, collecting and annotating such datasets can be time-consuming, expensive, and challenging. Accurate labeling of training data is crucial for effectively training AI systems.

References

[1] Agnese Augello and Giovanni Pilato. An annotated corpus of stories and gestures for a robotic storyteller. 2019. 2

[2] Michihiro Shimada Fumitaka Yamaoka Hiroshi Ishiguro Chao Shi, Takayuki Kanda and Norihiro Hagita. development of communicative behaviors in social robots. in proceedings of the ieee/rsj international conference on intelligent robots and systems. 2010. 2

[3] Karl F. MacDorman Daisuke Matsui, Takashi Minato and Hiroshi Ishiguro. Generating natural motion in an android by mapping human motion. 2005. 1

[4] Rodolphe Gelin Emmanuel Pot, Jérôme Monceaux and Bruno Maisonnier. Choregraphe: A graphical tool for humanoid robot programming. 2009. 1

[5] Michelle Annett Michael Wang Daniel Wigdor Haijun Xia, Michael Glueck. Iteratively designing gesture vocabularies: A survey and analysis of best practices in the hci literature. 2022. 3

[6] Stefan Kopp Maha Salem and Frank Joublin. Generating finely synchronized gesture and speech for humanoid robots: A closed-loop approach. 2013. 2

[7] Yibing Nan Kai Wang Hao Chen Minjie Hua, Fuyuan Shi and Shiguo Lian. Towards more realistic human-robot conversation: A seq2seq-based body gesture interaction system. 2019. 1

[8] Gentiane Venture and Dana Kulić. Robot expressive motions: A survey of generation and evaluation methods. In *ACM Transactions on Graphics (TOG)*, 2019. 2

[9] Pengcheng Luo Victor Ng-Thow-Hing and Sandra Okita. Synchronized gesture and speech production for humanoid robots. 2010. 2

[10] Juan Pabl Wachs. Vision-based hand-gesture applications. 2011. 3