Augmenting the Communication Naturalness via A 3D Audio-Visual Virtual Agent for Collaborative Robots

Rui Li
Department of Computer Science
Montclair State University
Montclair, NJ, USA
liru@montclair.edu

Weitian Wang
Department of Computer Science
Montclair State University
Montclair, NJ, USA
wangw@montclair.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671664

***Abstract*—In human-robot collaboration, current widely used human-robot communication is mainly based on audio and haptic mediums. However, this kind of communication is stiff and mechanical. Inspired by human-human communication in which vision and hearing contribute over 88% for human perception, we propose a knowledge-driven audio-visual virtual agent system, which allows collaborative robots to present its knowledge and feelings in a human-like way. During the collaboration training process, the virtual agent will build its assembly knowledge of how to work with the co-worker based on inverse reinforcement learning. To deploy a co-assembly task with its human partner, the virtual agent will also be able to produce assembly knowledge-based responses, which include knowledge-driven speech and speech synchronized facial animations. By leveraging the proposed knowledge-driven virtual agent, the collaborative robot not only can fulfill the co-assembly task but also can communicate with human partner in a more natural way.**

Keywords—communication naturalness, virtual agent, human-robot interaction, collaborative robots

I. INTRODUCTION

Current robots are aiming to improve the work efficiency of human beings [1-4]. In manufacturing contexts, they can be designed and programmed to help human workers finish assembly tasks. Even though the robots have improved the productivity of manufacturers and shared the burden of the human workers, their stiff and mechanical communication manners trigger the tiredness and tension of human workers during assembly processes. This is one of the main reasons of why most of current robots in manufacturing environments are still caged to work on tasks instead of standing beside humans.

In recent years, some researchers focused on human-robot communication via natural language processing [5, 6]. However, this natural language-based communication method only attempts to control the robot in speech and receive feedback from the robot through audio, which are not user-friendly for general workers. Some studies were conducted to create humanoid robots to enhance human-robot interaction [7, 8]. For example, the Sophia robot from Hanson Robotics [9]. However, the uncanny valley effect makes these robots not suitable for working with non-robotics-professional people in manufacturing environments. Digital human is an alternate way to enhance the human-robot communication [10]. The Baxter Robot [11] is a user-friendly companion, which has an animated face on its LCD screen displaying expression during works. However, this simplified 2D animated face, which only includes the eyebrows and two eyes, is very limited for realizing a realistic and natural interaction process between the human and the robot when they are working together. The audio-visual humanoid agent is a 3D virtual agent, which has potential advantages to augment the quality of human-robot interaction for co-assembly tasks in manufacturing contexts [12].

To improve humans' subjective feelings in robots [13], this work aims to augment the communication naturalness via developing an audio-visual virtual agent [14] between humans and robots. We propose a 3D user-friendly audio-visual virtual agent, which has a capacity of producing the knowledge-driven speech and speech synchronized facial expressions for robots during the human-robot collaboration process. The human intentions in the collaborative tasks can be understood and learned by robots based on the inverse reinforcement learning algorithms [15]. Afterwards, the virtual agent is able to actively interact with the human and control the robot using its learned strategy. By leveraging the proposed 3D multimodal virtual interface, the human may interact with the robot naturally and visually to accomplish collaborative tasks.

The contributions of this work are summarized as follows: (1) A 3D knowledge-driven audio-visual virtual agent is proposed for robots to interact with human partners in collaborative tasks. (2) The virtual agent system could augment the communication naturalness via leveraging not only speech feedback but also visual feedback of robots.

II. MODELING METHODOLOGY

A. Assembly Knowledge

The facial expressions and speech data for driving the virtual agent are acquired from a human speaker through a camera and a microphone. The human speaker is required to speak assembly instructions that are related to the collaborative task. The speech synchronized facial expressions of the human speaker are extracted by transfer learning based facial tracking method [16]. This collected data will be further used to create the facial expressions of the virtual agent.

In human-robot collaborative tasks, the human worker's assembly actions are collected through a camera mounted on the collaborative robot, a sensory system worn on the human's forearm, and a natural language processing system. The acquired human action information is then parameterized as assembly knowledge for the virtual agent.

B. Robot Visual Expression Model

To visualize the speech synchronized facial expressions for the virtual agent, we apply a parametric facial model [14] in this study. Control points $\rho(t)$ are defined for driving the movements of the facial model:

$$f(t) = f(t-\Delta t) + \delta(t-\Delta t)\cdot\cos\left(\frac{2\cdot\pi\cdot d_i(t-\Delta t)}{\omega}\right) \quad (1)$$

where $f(t)$ denotes the shape of human face, the width of the mouth region can be calculated by ω, $\delta(t) = \rho(t) - \rho(t-\Delta t)$ indicates the changes of corresponding control points over the time, and d_i indicates the distance between the i^{th} facial region and the control points. The further distance indicates the less deformation impact on the facial region.

C. Eye Model

The emotionally expressive eye movement is important for human-robot communication. Hence an eye model is utilized to realize a natural multimodal communication. The movements of eyeballs and eyelids are considered for the simulation of eye movements. For the movements of eyeballs, three parameters (amplitude, duration, and direction) are used for controlling saccades. The movements of eyelid include the lid saccades and the blink. The lid saccades are controlled by three parameters: the lid position, the lid amplitude, and the lid saccades duration. The blink is controlled by three parameters: the blink amplitude, the blink duration, and the blink rate. For the synthesis of the expressive eye movement, six emotions (happy, surprise, anger, sadness, care, and fear) are simulated.

D. Facial Model

A 3D facial mesh, which includes 6294 points and 11800 triangles, is utilized in this work. The facial texture comes from a front view photo of a female volunteer. In order to simulate the facial animations, a parametric facial model [14] is built. Through tracking and analyzing the lips movements in large amount of human speaking video, six parameters are obtained to control the motion of middle points on the upper and lower lips in 3D spaces. The rest area around the lips is controlled based on the distances between a corresponding mesh point and the middle points on the two lips. In order to realize the natural and smooth deformation around the lips, a cosine function is employed when updating the positions of the mesh points to realize the lips deformation.

III. Preliminary Results

A. Experimental Setup

A human-robot collaborative task is designed to verify the efficiency of the proposed virtual agent in real-world smart manufacturing contexts. In this experiment, the collaborative robot and its human partner are required to co-assemble a vehicle model. The actions of the robot are classified as picking up the parts, placing the parts, delivering the parts to the human, and receiving the parts from the human. The actions of the human include receiving the parts from the robot, assembling the parts on the vehicle, and delivering the parts to the robot. In our experiment, the collaborative robot and the virtual agent are viewed as a whole robot system. For the following experimental descriptions, we will use the robot to denote this system.

B. Virtual Agent based Human-Robot Interaction Framework

The developed framework of human-robot interaction via the 3D audio-visual virtual agent in collaborative tasks is shown in Fig. 1. It includes (1) communication between the human worker and the virtual agent, (2) control commands for the robot, and (3) collaboration between the human and the robot.

For the human-robot communication, the robot understands and learns human intentions via natural language and gesture information characterized by the camera and the wearable sensory system. Meanwhile, it can also express its own knowledge and feelings to the human. The human-robot communication can be enhanced by the audio-visual information performed by the robot through the virtual agent. To express knowledge, the robot will be able to use natural language and speech synchronized facial animations to respond humans. The responses will be in the form of not only the speech but also the speech synchronized facial expressions.

In addition to fulfill the assembly task with its human partner, the robot is able to express its feelings based on its own judgement of the current situation. Six emotions, including happy, surprise, anger, sadness, care, and fear, are used to augment the robot's feeling expressions. For example, when the human becomes tired, the virtual agent will produce a facial animation with caring. Meanwhile, the collaborative robot arm will slow down its actions. When the human partner is high-energy, the virtual agent will express happiness and speed up the robot arm. When the human partner performs some dangerous operations such as a long-time placing of hands on the center of assembly table, the virtual agent will express fear, stop the robot arm, and remind the human to move hands back.

In the human-robot collaboration process, the corresponding control commands will be generated to control the robot to collaborate with its human partner to complete the collaborative tasks such as picking up the parts, placing the parts, or delivering the parts to the human.

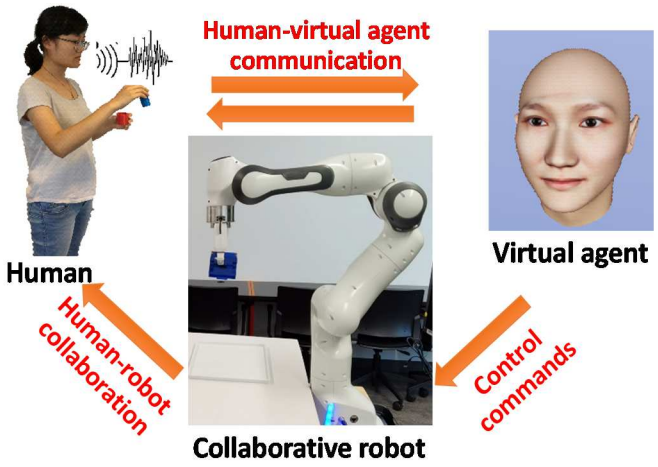

Fig. 1. The framework of human-robot interaction via the proposed audio-visual virtual agent in collaborative tasks.

IV. Conclusions and Future work

We proposed a 3D audio-visual virtual agent system to enhance the communication naturalness for the human and the robot in collaborative tasks. The virtual agent has a capacity of producing the knowledge-driven speech and speech synchronized facial expressions during the human-robot collaboration process. By leveraging the proposed 3D virtual agent, the robot can communicate with its co-worker in a human-like way to finish collaborative tasks. In addition to respond to human requests by natural language and speech synchronized facial animations, the robot is able to express its own feelings (six emotions) in the working process. Future work of this study will focus on implementing the proposed virtual agent to real-world complicated human-robot collaborative tasks and evaluating human factors (e.g., comfort and trust levels of the human) with/without the virtual agent.

Acknowledgment

This work was supported in part by the National Science Foundation under Grant CNS-2104742 and in part by the National Science Foundation under Grant CNS-2117308.

References

[1] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics,* vol. 55, pp. 248-266, 2018.

[2] E. Matheson, R. Minto, E. G. Zampieri, M. Faccio, and G. Rosati, "Human–Robot Collaboration in Manufacturing Applications: A Review," *Robotics,* vol. 8, no. 4, p. 100, 2019.

[3] W. Wang, R. Li, Z. M. Diekel, Y. Chen, Z. Zhang, and Y. Jia, "Controlling Object Hand-Over in Human–Robot Collaboration Via Natural Wearable Sensing," *IEEE Transactions on Human-Machine Systems,* vol. 49, no. 1, pp. 59-71, 2019.

[4] W. Wang, R. Li, Y. Chen, Y. Sun, and Y. Jia, "Predicting Human Intentions in Human-Robot Hand-Over Tasks Through Multimodal Learning," *IEEE Transactions on Automation Science and Engineering,* pp. 1-15, 2021, doi: 10.1109/TASE.2021.3074873.

[5] S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein, "Mobile robot programming using natural language," *Robotics and Autonomous Systems,* vol. 38, no. 3, pp. 171-181, 2002.

[6] S. A. Tellex *et al.*, "Understanding natural language commands for robotic navigation and mobile manipulation," 2011.

[7] M. Hirose and K. Ogawa, "Honda humanoid robots development," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 365, no. 1850, pp. 11-19, 2007.

[8] B. Adams, C. Breazeal, R. A. Brooks, and B. Scassellati, "Humanoid robots: A new kind of tool," *IEEE Intelligent Systems and Their Applications,* vol. 15, no. 4, pp. 25-31, 2000.

[9] J. Retto, "Sophia, first citizen robot of the world," *ResearchGate, URL: https://www. researchgate. net,* 2017.

[10] D. B. Chaffin and C. Nelson, *Digital human modeling for vehicle and workplace design*. Society of Automotive Engineers Warrendale, PA, 2001.

[11] S. Cremer, L. Mastromoro, and D. O. Popa, "On the performance of the Baxter research robot," in *2016 IEEE international symposium on assembly and manufacturing (ISAM)*, 2016: IEEE, pp. 106-111.

[12] H. Tang, Y. Fu, J. Tu, M. Hasegawa-Johnson, and T. S. Huang, "Humanoid audio–visual avatar with emotive text-to-speech synthesis," *IEEE Transactions on multimedia,* vol. 10, no. 6, pp. 969-981, 2008.

[13] C. Hannum, R. Li, and W. Wang, "Trust or Not?: A Computational Robot-Trusting-Human Model for Human-Robot Collaborative Tasks," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 5689-5691.

[14] R. Li and J. Yu, "An audio-visual 3D virtual articulation system for visual speech synthesis," in *2017 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE)*, 2017: IEEE, pp. 1-6.

[15] W. Wang, R. Li, Y. Chen, Z. M. Diekel, and Y. Jia, "Facilitating Human–Robot Collaborative Tasks by Teaching-Learning-Collaboration From Human Demonstrations," *IEEE Transactions on Automation Science and Engineering,* vol. 16, no. 2, pp. 640-653, 2018.

[16] H. Diamantopoulos and W. Wang, "Accommodating and Assisting Human Partners in Human-Robot Collaborative Tasks through Emotion Understanding," in *2021 International Conference on Mechanical and Aerospace Engineering (ICMAE)*, 2021: IEEE, pp. 523-528.