Adaptive Gesture Extraction and Imitation for Human-Robot Interaction

Konstantinos Theofilis, Chrystopher Nehaniv, and **Kerstin Dautenhahn**
Adaptive Systems Research Group
University of Hertfordshire
Hatfield, AL10 9AB
United Kingdom

Background and related work

Gestures can be characterized as body movements that are used to convey information from one person to another (Vaananen and Bohm 1993). As such, they are a crucial part of human-human interaction. Gestures are even more important in the early stages of human development, where the lack of speech promotes the gestures as the main communicative mode. However, even for humans, the context of a gesture can vary considerably depending on a wide range of factors. Even cultural differences can completely change the meaning of a gesture.

The importance of gestures in human communication makes them crucial in Human-Robot Interaction (HRI). However, successful recognition, grounding and reproduction of gestures remain a major challenge in robotics. Apart from visual realism, successful use of gestures by robots must also adhere to behavioural realism. Behavioural realism is more important than visual realism and a wrongly used gesture can cancel the interaction loop needed for a successful HRI. Early approaches were using only basic predefined gestures. Machine vision could not provide accurate tracking of the human body without cumbersome and expensive sensors and great computational demands. Later, the availability of cheap RGB-Depth sensors provided the means for robots to easily and accurately track humans and recognise some basic gestures, but still depending on a predefined mapping of the gesture to be recognised and reproduced.

Newer approaches use training algorithms with manually annotated data of human gestures as training data (Kipp 2003). While these approaches increase the efficiency of the gesture recognition, they remain inflexible regarding the context of the gesture. Other efforts for solving the problem try to bypass the grounding issue by binding the information load of a gesture to speech (Cassell 1998). Recognising what the human is saying during the gesture is used to classify the gesture (Salem et al. 2012), (Chiu and Marsella 2014).While successful in certain scenarios, the obvious drawback is that the gesture grounding issue is delegated to context extraction from speech recognition, itself also a major challenge (Levine, Theobalt, and Koltun 2009). Also, these approaches fail to take into account what happens when speech is not available or practical to be used during the interaction, making them inapplicable in many HRI scenarios.

Proposed approach

The dynamic and adaptive approach proposed here is based on contextually important points in time. These can be predefined or automatically recognized. The idea is that these important points of the interaction (e.g., the end of a turn in turn-taking) carry their own information load, so gestures that happen from one participant of the interaction around these points are used as means to transfer this information to the other participant.

Also, in HRI, these key points of the interaction timeline have to identified anyway in most cases, regardless of the need for gesture extraction and imitation. Therefore, an approach based on them does not significantly add to the cognitive and computation load of the robotic system.

A diagram showing a simple application of the approach is shown in Fig. 1. The gesture G1 performed by the human at the end of the first turn (during which the human is the actor), is extracted using RGB-D sensors and used by the robot at the end of turn 2 (with robot as the actor). The gesture G1 is internally stored and classified by the robot as a gesture that signals the end of a turn.

Another secondary direction that is researched using this approach, is the extraction of gestures that can be classified as unconscious mannerisms not directly mapped to information (Nehaniv et al. 2005). E.g., in Fig. 1, while the robot is the actor, the human is performing a gesture G2. The robot is storing this gesture to be performed later when the human is the actor. While this type of gestures may not directly aid the information exchange, they have the potential to increase the behavioural realism of the robot.

System description and implementation

Our system was implemented in order to be used in HRI studies regarding communicative imitation. The robotic platform used was the iCub humanoid robot (*http://www.icub.org*). The use of the iCub with its numerous degrees of freedom contributed to the visual realism of the produced gestures however, good results could be achieved using other humanoid robots.

The setup with which the system was tested, was a simple interactive game between the iCub robot and a human. The human was sitting in front of a table that had 4 large coloured buttons that produce musical notes. The human was demonstrating a tune to the robot and then the robot could play it back or demonstrate its own tunes. In some cases, the human would choose to demonstrate the tune more than one time before the robot would try to reproduce it. In the setup used, the context-sensitive time points were the pressing of the buttons and the end of each demonstration of the tune.

The gesture extraction and reproduction system is comprised of several software modules. Its structure can be seen in Fig. 2. A module is constantly tracking and storing the human's skeleton position and orientation for 15 joints receiving data from a Xtion Pro RGB-D sensor using the OpenNI2 API and the NiTE2 middleware for skeleton tracking. The second sensor is used for tracking the five fingers of the dominant hand of the human, using the FORTH middleware. The finger tracking was used mostly for evaluating the possibility of integrating such data into the gesture recognition and reproduction but focus was mainly on the arm imitation.

A *Watchdog* module is tracking important events and their timing. In this case, the important events were the pressings of the buttons and the end of a sequence played by the human. This module was needed for the purposes of our study regardless of the gesture system.

The *Gesture extractor* module is notified from the watchdog that a keypoint is identified and retrieves the human's arm movements around that time point. Heuristic rules are applied depending on the scenario. In the test case, the movements *after* the button press are extracted.

The *Safeguards filter* module receives the extracted gesture and checks if it conforms to rules that are necessary for the safety of the robot. For example, if the imitation of the gesture would mean that part of the robot's arm would hit the table, the gesture is limited in the direction that would cause the breakage. Then, the gesture and its associated metadata are stored in a database.

When the *Gesture chooser* module is informed that a time point for a gesture reproduction is approaching, it retrieves the associated gesture, and sends it to the robot's kinematics module for reproduction. If there is more than one gesture with the same context stored (e.g., if there is more than one gesture marked as end of turn), a similarity check is performed and further heuristics are used to choose the one to be reproduced.

Finally, the gesture is reproduced using velocity control. The advantage of using velocity control is that the absolute start and end positions of the human's arms are not important and only the relative orientation of the arm's joints are reproduced.

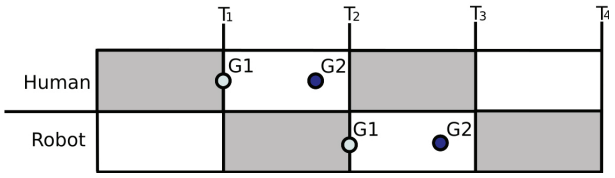

Figure 1: Sample sequence diagram. G1 and G2 represent gestures, T1,T2,T3,T4 represent turns 1-4. Gray areas represent the active actor (human or robot) for that turn.

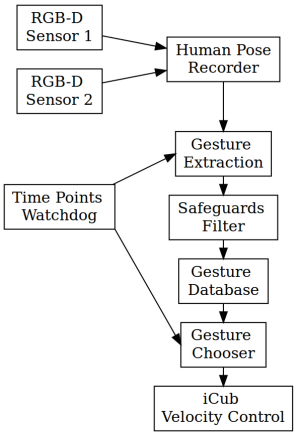

Figure 2: Continuous gesture extraction and reproduction system.

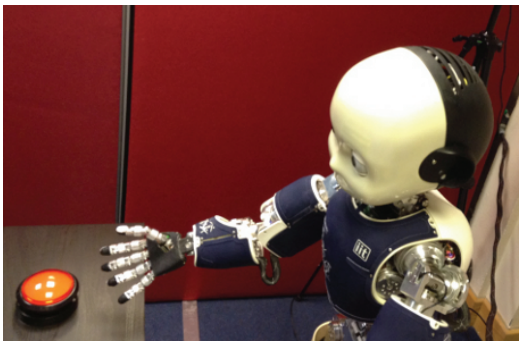

Figure 3: The iCub robot performing a suggestive gesture to signal the end of its turn. The gesture was the same that the human participant performed at the end of the previous turn.

Conclusion

The approach presented has the advantage that extracts most of its necessary context-related information from systems that would have to be implemented, regardless of the need for gesture extraction. At the same time, it avoids the need for time-consuming training of the system. This model lends itself well to the deferred nature of gesture imitation in turn-taking scenarios in HRI. The main overhead has to do with the safeguards and filters that have to be specific to robots and setups.

While further work is needed to improve parts of the system, the current implementation allows for gesture extraction and reproduction that appears to be visually and behaviourally realistic and at the same time is dynamic and adaptive to individual users without the burden of speech-based approaches and the inflexibility of predefined systems.

References

Cassell, J. 1998. *Computer Vision in Human-Machine Interaction*. Cambridge University. chapter A framework for gesture generation and intepretation.

Chiu, C.-C., and Marsella, S. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, 781–788. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Kipp, M. 2003. *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*.

Levine, S.; Theobalt, C.; and Koltun, V. 2009. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.* 28(5):172:1–172:10.

Nehaniv, C.; Dautenhahn, K.; Kubacki, J.; Haegele, M.; Parlitz, C.; and Alami, R. 2005,. A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication* 371–377.

Salem, M.; Kopp, S.; Wachsmuth, I.; Rohlfing, K.; and Joublin, F. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics, Special Issue on Expectations, Intentions, and Actions* 4(2):201–217.

Vaananen, K., and Bohm, K. 1993. *Virtual Reality Systems*. Academic Press, Ltd. chapter Gesture-driven interaction as a human factor in virtual environments - An approach with neural networks.