Generative Methods for Object-Based Gestures for Socially Interactive Robots

Eric C. Deng
University of Southern California
3710 McClintock Avenue
Los Angeles, California 90089
denge@usc.edu

Maja J. Matarić
University of Southern California
3710 McClintock Avenue
Los Angeles, California 90089
mataric@usc.edu

Abstract

Socially interactive robots leverage embodied cues to augment more traditional modes of communication, such as speech, text, and screen-based interactions, to engage in rich interactions with users. We aim to endow these embodied agents with the ability to perform object-based gestures by formalizing a process for generating such gestures and providing tools for efficiently developing controllers for them. We present a proposed pipeline for generating *object-based gestures*, describe tools we have and continue to develop for automating modules of this pipeline, and discuss challenges for future work in gesture generation.

Introduction

Gesture is a key part of human embodied communication. As human-robot interaction (HRI) seeks to leverage natural human-human communication, gestures can be highly effective in improving user experience with artificial agents. In this paper, we focus on efficient methods for developing object-based gestures designed to communicate ideas tied directly to semantics in associated speech or text. Using traditional models of objects as the input for generating embodied behaviors, we propose a flexible system for autonomously generating gestures that are then mapped to different robot embodiments. This paper is organized as follows: we discuss existing work in relevant fields of research, our proposed pipeline for generating speech-paired, object-based gestures, descriptions of the different modules within that pipeline, and current and ongoing development of two modules within the proposed pipeline.

Background

Generating robot gesture involves a collection of key components, including grounding speech and text-based interactions, selecting appropriate objects to gesture, generating optimized gestures from those objects, and, for generality, mapping those gestures to different robots.

Semantic Embedding in Traditional Media Forms

In explicitly communicative human-agent interaction, embodied gestures can be used to improve the agent's ability to refer to abstract or specific objects. For disambiguation, embodied gestures are often most effective when paired with speech or text, and therefore need to be grounded in cues of those modalities.

Work in natural language processing (NLP) and natural language understanding (NLU) has explored *semantic embedding* in text, seeking understand of the meaning of segments of language and representing them to be compared and related to other forms of media, such as 2D images and video (Frome et al., 2013). Research in semantic embedding has produced methods for automatically generating non-templated descriptions of 2D images and videos using convolutional neural networks (CNNs) over content and aligned using multi-modal recurrent neural networks (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015). This work is likely to be expanded to more media forms, such as 3D models or 3D animations, and will be even more applicable and useful for our application. Using semantic understanding applied to 2D and 3D images and video can be used to search and select the most relevant objects to be gesturally expressed, concurrently with relevant segments of speech or text (Jurafsky, 2000). Semantic embedding is critical to our approach as it allows us to have shared terminology to relate traditional media forms with the gestures we aim to generate.

Robotics and Embodied Communication in HRI

In HRI, researchers have studied embodied gestures to improve the efficiency and quality of robot communication, designing and evaluating cues such as gaze behavior (Andrist et al., 2012), legible motion (Holladay, Dragan, and Srinivasa, 2014; Nikolaidis, Dragan, and Srinivasa, 2016), and proxemics (Takayama and Pantofaru, 2009). To augment speech-based interactions, we aim to develop a method for generating object-based gestures, which tend to be in one of two gesture categories: *iconic* or *metaphoric* McNeill (2008). Those two gesture categories may account for as much as 76 % of gestures used in speaking / narrative tasks (Takeuchi et al., 2017). Research to date has already shown how embodied gestures can generally improve overall interaction capabilities of robots and how important object-based gestures are in the context of embodied gestures as a whole (Huang and Mutlu, 2013; Takeuchi et al., 2017).

Most embodied gestures, including object-based ones, are preprogrammed (Sugiyama et al., 2007; Ido et al., 2006)

or based on motion capture data for specific agent embodiments and kinematic configurations (Matsui et al., 2005). In contrast, we build on our prior work in Deng and Mataric (2017) that outlines an adaptation of the space-matter manipulation technique from the art of mime to produce gestures about the physical shape and properties of objects.

Generative Models and GANs

One scalable approach for creating objects that can serve as a basis for embodied gestures is to use generative models Jaakkola and Haussler (1999) and generative adversarial networks (GANs) Goodfellow et al. (2014). These networks use a discriminator and a generator and can build realistic examples of the content upon which the system is trained. The discriminator network is a system that acts as a traditional classifier that works to classify different instantiations of a certain form (image, video, etc.) into categories based on some set of training data. The generator network is a system that, given a category, creates what it believes is an accurate instantiation of what an object of that categorization *could* be. By connecting these two networks, we can have them "compete" against each other and incrementally improve the performance of the overall system by feeding the examples generated by the generator into the discriminator network and updating the discriminator network model based on the known classification of the object being generated. They have been used in relevant applications such as text-to-image generation Reed et al. (2016). Given enough data properly contextualized in social interaction, GAN architectures can be used to generate embodied gestures.

Object-Based Gesture Generation

The goal of this work is to develop methods for efficiently generating object-based iconic and metaphoric gestures for socially interactive robots. Although our focus is on gesture use in robotics, the approach can also be applied to virtual agents and animation.

Our system looks ahead in its planned speech while generating multi-modal behaviors and selects and generates appropriate behaviors. Figure 1 outlines our characterization of how an interactive robot can generate object-inspired, multimodal interaction with utterances as the input, broken down into four primary modules. The first module consists of the *natural language system* that takes in the robot's utterance as input and searches for a set of potentially relevant objects to gesture leveraging an existing NLP method. In the example shown, the robot says "This is a ball", the system extracts "ball" as the subject noun, and searches for different instances of that noun/object that can be used as references for generating a gesture. From this set of categories of objects (i.e., different types of balls) we select one object, in this case *basketball*, based on the greater context of the interaction, or randomly in the absence of contextual clues. The second module, *reference object selection*, then selects the best example of the object from the avalabile set of examples (which may be in various forms, such as 2D images, 3D models, animations, etc.) that is then used as the reference object for generating the gesture. The selected example in its

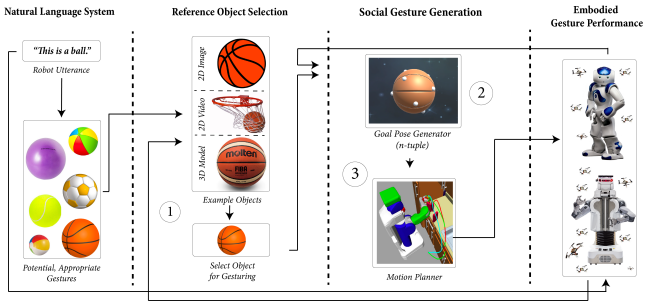

Figure 1: The gesture generation pipeline from speech-based input to object-based, end-effector gestures.

native form is passed into the third module, *social gesture generation*, and parsed into a set of 6 degrees-of-freedom (DoF) poses for use as potential goal poses for the gesture. To generate the gesture, the module parses the model into points distributed along the edges or surfaces of the object, generating a trajectory. Motion planning techniques can then be used to produce the goal trajectory with the given robot manipulator end-effector. The trajectory is executed concurrently with the input utterance, as seen in the fourth module, *embodied gesture performance*. Various optimizations to this pipeline are possible. We focus on *reference object selection* and *social gesture generation* and have designed and built a tool on the Unity game engine and Robot Operating System (ROS) Quigley et al. (2009) to streamline those two modules.

Reference Object Selection

The second module in Figure 1 is concerned with selecting an example of an object to be used as the model for the gesture to be generated. A query about the object in a Web search could return a variety of examples, such as 2D images, 3D models, videos, etc. The ranking of the search results will not be optimized for salience and identifiability in gesture form.

Our system is a small-scale prototype of what could be a large-scale, cloud-based gesture generation pipeline with access to databases of different instances of object types, such as Google. This is because we are contributing to the *approach* for converting traditional media forms into robot gestures. Our system currently has an "object database" of around a dozen 3D models of basic objects like balls, boxes, and pyramids. Because the prototype database only includes one example of each type of object, there is no need for sorting of objects. In general, however, the objects could be ranked using a social metric relevant for object-based gestures. Potential quantifiable factors in object salience include:

1. *Interaction Context*–where, when, and with whom the robot is interacting. For example, if the robot is to be placed in a public space and meant to interact with many people at once, a 3D model may be more appropriate as it is more salient from multiple perspectives.
2. *Gesture Strategy*–object form vs. object use. Iconic ges-

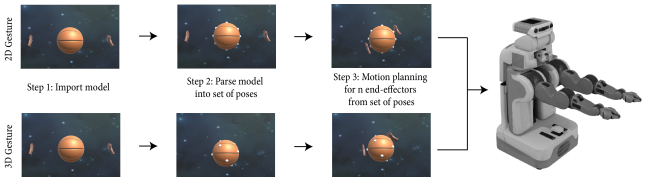

Figure 2: Current methods for generating 2D and 3D gestures based on 3D models built in Unity and performed via ROS and RViz.

tures sometimes use objects as tools for demonstrating ideas being communicated. Rather than gesturing the physical outline of tools, these iconic gestures may show tool use instead. For example, a hammer is expressed with a hammering gesture. Animations and videos can be used to generate such gestures.

3. *Speech Context*–how important a gesture is in relation to the complete expressive speech being generated by the robot. This directly relates to how long and involved the gesture is (the less important, the less "interaction resources" should be used) and therefore constrains how complex the example object should be. The more complex, the longer it takes to complete a gesture (to minimize aliasing), influencing the selection of the example objects.
4. *Kinematic Configuration*–the physical properties of the robot. The kinematic constraints and other motion planning factors can help to select object examples with more viable expressive trajectories.

Establishing a mapping of features of traditional representations of objects to salience in end-effector gestures is a key step. Once made, object sorting is greatly simplified, and object can even be generated using techniques such as those presented in (Reed et al., 2016). GANs could allow for bypassing defining this representation-to-salience mapping if enough labeled data are available of human responses to different gestures generated from a variety of object representations. The results from text-to-image generation using GANs indicated that, with the right type of data, object representations can be generated for conversion into end-effector controllers directly from text (Figure 1).

Social Gesture Generation

The previous module output a traditional representation of the selected object. In this module, the representation is converted into "robot-gesturable" forms, and then those are used as input to motion planning.

Segmenting traditional media forms to a set of 6 DoF poses is done by converting the existing form into a number of points that can be used as potential goal poses for generating an end-effector trajectory. The number of points and the density distribution of the points over an edge or surface of 2D or 3D models, respectively, are determined by the complexity of the original model and the number of end-effectors available to the robot. Our system currently distributes the poses evenly across the surface or edge without considering the topology but we are continuing to explore how topology may affect salience for object-based gestures. Converting these points into goal poses involves solving for the Cartesian coordinates, transforming the image point into the end-effector 3D space (scaling and shifting according), and generating Euler angles by finding the normal at those coordinates.

Given a set of potential goal poses, the system groups those into a set of sequential n-tuples, with n being the number of end-effectors available. Generating a trajectory through this set of n-tuples, allows each end-effector to perform the gesture and makes the system platform-agnostic until the final planning stage. The system currently generates these n-tuples based on average Euclidean distance (non-optimally for feasibility) and users can manually select and modify the sets and the order for the robot to perform. After the n-tuples have been generated and ordered, the goal poses are fed into the motion planner for the given robot and then performed. Figure 2 shows the approach for both 2D and 3D object models.

Our generation pipeline system implementation currently relies on the user to edit and select the most effective sets of poses. To avoid trajectory collisions, we sequentially plan for each end-effector in the system and use the existing trajectories as obstacles. Although this may not be the most efficient method for planning currently it is robust and allows us to quickly evaluate endpoint selection algorithms with are more important to salience than motion planning. Our goal is toward automatically optimizing for factors such as these:

1. *Viewpoints*–the perspective that interaction partner(s) are taking during the gesturing. This can be used to transform the example object appropriately (e.g., by rotating it so that the most identifiable perspective is facing the user) or to generate more legible gestures relative to for occlusions and depth uncertainty.
2. *Time of overall gesture*–similar to the factor from the previous section on object selection in that once the object is selected, the gesture needs to be selected so that it can be completed within the predetermined length of time. Our current system uses a single point to represent each n-tuple (as that is how the n-tuples are generated) and to generate the trajectory uses the bisection method in reference to the distances between those reference points. We aim to evaluate other methods.

Conclusion and Future Work

This paper has described work that aims to generate object-based gestures with the goal of creating more communicative and engaging socially interactive robots. Such gestures make up a large portion of human gestures during narration and can therefore be effective tools for improving embodied communication between robots and people. In this work we presented a gesture generation pipeline from speech, discussed the two implemented modules of that pipeline, and introduced potential improvements.

This work can be described in a 20-minute presentation on our motivations for this work, current system design, upcoming user study designs, video examples of the system,

and gestures produced in simulation and by real robots. By March 2018, we will also have a tool that can be used for an interactive demonstration–allowing users to select models, parse them, generate gestures, and "play back" the gestures on a simulated PR2 robot. This demo may also be paired with a rudimentary voice-based interaction that will allow users to submit input via voice commands as well.

We plan to complete this tool and evaluate different search methods and heuristics for *social gesture generation*, building and testing the salience of the output behaviors both with Mechanical Turk participants, and with co-located viewers with convenience populations of college students. We will also explore predicting gesture salience from object models to be used for optimization of the *reference object selection* module, again using both online and in-person experiments with gestures generated with the "optimized" behaviors from the updated *social gesture generation* module.

References

Andrist, S.; Pejsa, T.; Mutlu, B.; and Gleicher, M. 2012. Designing effective gaze mechanisms for virtual agents. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 705–714. ACM.

Deng, E., and Mataric, M. J. 2017. Mime-inspired behaviors in minimal social robots. In *ACM CHI Workshop on What Actors can Teach Robots*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Holladay, R. M.; Dragan, A. D.; and Srinivasa, S. S. 2014. Legible robot pointing. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, 217–223. IEEE.

Huang, C.-M., and Mutlu, B. 2013. Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems*, 57–64.

Ido, J.; Matsumoto, Y.; Ogasawara, T.; and Nisimura, R. 2006. Humanoid with interaction ability using vision and speech information. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, 1316–1321. IEEE.

Jaakkola, T., and Haussler, D. 1999. Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems*, 487–493.

Jurafsky, D. 2000. *Speech & language processing*. Pearson Education India.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Matsui, D.; Minato, T.; MacDorman, K. F.; and Ishiguro, H. 2005. Generating natural motion in an android by mapping human motion. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, 3301–3308. IEEE.

McNeill, D. 2008. *Gesture and thought*. University of Chicago press.

Nikolaidis, S.; Dragan, A.; and Srinivasa, S. 2016. Viewpoints-based legibility optimization. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*, 271–278. IEEE.

Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; and Ng, A. Y. 2009. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, 5. Kobe.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Sugiyama, O.; Kanda, T.; Imai, M.; Ishiguro, H.; and Hagita, N. 2007. Natural deictic communication with humanoid robots. In *Intelligent robots and systems, 2007. IROS 2007. IEEE/RSJ international conference on*, 1441–1448. IEEE.

Takayama, L., and Pantofaru, C. 2009. Influences on proxemic behaviors in human-robot interaction. In *Intelligent robots and systems, 2009. IROS 2009. IEEE/RSJ international conference on*, 5495–5502. IEEE.

Takeuchi, K.; Kubota, S.; Suzuki, K.; Hasegawa, D.; and Sakuta, H. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction*, 198–202. Springer.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Authors

Eric Deng is a 4th year BS/MS electrical engineering and mechanical engineering Stamps Scholar emphasizing in robotics at USC. As a Richardson Research Fellow, Eric's current research interests include nonverbal gestures, joint attention, and computational embodiment design. He has work experience at IDEO, Facebook, Botkins Robotics, VNTANA, and is currently a Product Engineering Associate at Embodied Inc.

Maja J Matarić is professor and Chan Soon-Shiong chair in Computer Science Department, Neuroscience Program, and the Department of Pediatrics at the University of Southern California. Her Interaction Lab's research into socially assistive robotics is aimed at endowing robots with the ability to help people through individual non-contact assistance in convalescence, rehabilitation, training, and education.