Teaching Arm and Head Gestures to a Humanoid Robot through Interactive Demonstration and Spoken Instructions

Michael Connolly Brady
DFKI, Saarland Informatics Campus
66123, Saarbrücken, Germany
michael.brady@dfki.de

Han Du
DFKI, Saarland Informatics Campus
66123, Saarbrücken, Germany
han_h.du@dfki.de

Abstract

We describe work in progress for training a humanoid robot to produce iconic arm and head gestures as part of task-oriented dialogic interaction. This involves the development of a *multimodal dialogue manager* and corresponding system architecture for non-experts to 'program' the robot through speech and vision. Using this system, videos of gesture demonstrations are collected. Motor positions are extracted from the videos to specify motor trajectories, where collections of motor trajectories are used to produce robot gestures following a Gaussian mixtures approach. Concluding discussion considers how learned representations may be used for gesture recognition by the robot, and how the core system may mature into a robust system to address language grounding and semantic representation.

1 Introduction

A conventional way of programming robots to make iconic gestural movements is to animate movements as sequences of static motor positions. This method is slow and tedious and an easier method is sought. Ideally, people should be able to teach a robot how to make new gestures through visual demonstration and verbal instruction, as they might teach another person how to make a new arm and head gesture. Such a multimodal interactive approach is one of today's current challenges in robotics. Perhaps one reason that multimodal interaction with robots is problematic relates to the compartmentalization of research specialties. Speech engineers are generally not experts at computer vision and motor control. Likewise, robotics engineers and computer vision engineers tend to treat speech and language as a 'black box' problem best left to speech and language technologists. The result is that language, vision, and motor control tend to be segregated during software planning and implementation. It is left to the robot interaction engineer to cobble these segregated modalities together into a cohesive software framework. The broad aim of our project is to pragmatically address this challenge by developing a processing architecture where communicative information across modalities can be more integrated. Teaching a robot how to produce gestures through visual demonstration and spoken dialogue is a task that is well suited for addressing the challenge.

Robot Learning from Demonstration (LfD), sometimes also referred to as "robot programming by demonstration," "teaching by example," or "imitation learning" is an established approach for training robots through vision. As alluded to above however, one issue with LfD is that LfD practitioners generally fail to incorporate the power of verbal instruction, see (Ravichandar et al., 2020). We posit that with the relatively recent advent of Deep learning and related breakthroughs in computer vision, artificial speech recognition, and related technologies, the time is ripe to integrate natural verbal instruction with LfD.

LfD and training by example has a rich history and is a popular research area in modern robotics, for example see: (Calinon and Billard, 2007; Argall et al., 2009; Koenig et al., 2010; Calinon et al., 2010; Lee, 2017; Zhu and Hu, 2018; Ravichandar et al., 2020). LfD sidesteps more traditional and tedious methods of manually specifying motor control or where math and computer programing expertise is required. The essence of LfD is that robot movements may be acquired by having a person act out the movements to be learned (either through telepresence, kinetically, or visually), and transposing those movements into representations that a robot may use in combination with the robot's knowledge and internal processing to then produce the movement. It is important to note that LfD is not merely a 'record and replay'

Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), pages 95–101

Figure 1: 'VoxHead' 3D printed humanoid robot

technique. Generalization is required so that, for example, starting and ending positions of the movements are not pre-determined. Exact trajectories as well as amplitudes of movements may vary insofar as the task demands, and resulting movements should be robust in the face of changing environmental conditions and actuator imprecisions. For our present purpose, the idea is also to avoid exact monotonous repetitions, and to develop robust representations that may also be used for perceiving learned gestures.

Interacting with robots through natural language is another popular area of research. E.g. see: (Cantrell et al., 2010; She et al., 2014; Gemignani et al., 2015; Misra et al., 2018; Liu and Zhang, 2019; Kruijff-Korbayova et al., 2020). Perhaps the most popular domain for linguistic information transfer between people and robots is in giving travel or route instructions, such as in the spoken guidance of robotic wheelchairs, for a review see: (Williams and Scheutz, 2017).

It is important to note that speech communication also contains non-linguistic cues, both vocal (e.g. laughter, affect, tone) and non-vocal (e.g. gestures, eye gaze, face expressions, environmental context). For related review, see: (Mavridis, 2015; Devillers et al., 2020). In addition to the linguistic signal, these and related cues should be readily available for incorporation into interaction designs.

2 Method

The robot this work uses is "VoxHead," a 3D printed humanoid robot (Brady, 2016; Devillers et al., 2020). Figure 1 displays the robot. The robot serves as a life-sized and relatively low cost platform for interactive social robotics research. The robot has motors for mouth, eye cameras, and facial expressions. For the present work we do not concern ourselves with facial motors. Instead, focus is on general head, neck, and arm movements. In total there are sixteen degrees of freedom in the head, neck, and arms that we work with. Specifically we use: head tilt, head turn, neck tilt, neck turn, and for each arm: arm raise-lower, arm left-right, arm rotate, elbow bend, wrist rotate, and wrist bend. Hands with individual fingers or grippers are also not used here.

2.1 Control Architecture

Figure 2 depicts the general software plan. Sensory input to the robot is handled by a series of perception modules. A perception module may run on it's own mini-computer as e.g. an end-to-end DNN, or may run on a remote server, such as with an ASR engine. A countless number of perceptual processing modules may in theory be included, a few of which are portrayed here. For the present purpose of simplicity, only a speech-to-text ASR percep-

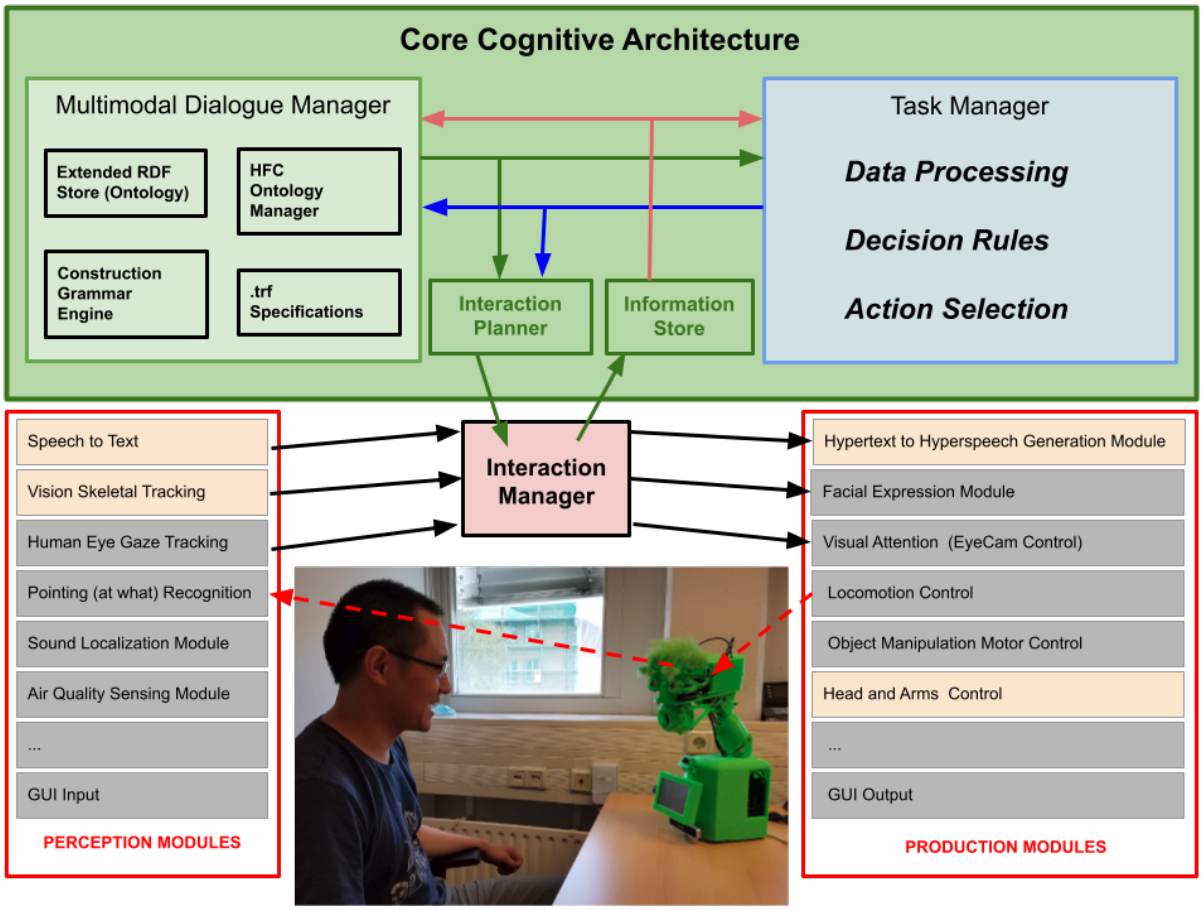

Figure 2: system architecture

tion module (Amazon Transcribe), and a skeletal tracking perceptual module (to be described in Section 2.2) are used. Input from these two sources is received by an *Interaction Manager (IM)*. The IM collects sensory input based on a control signal from the *Core Cognitive Architecture (CCA)*. Sensory input that is requested by the CCA feeds to an *Information Store (IS),* for cognitive processing. The IM also relays commands from the CCA to be executed by various production modules. Like with the perception modules, a countless number of production modules may be included, a few are portrayed, and for the present purpose only the two highlighted modules (speech synthesizer, and head and arms motors controller) are considered here.

The CCA is very much a work in progress. Skeletal tracking information is read by a *Task Manager (TM),* within the CCA for data processing (see Section 2.2), while linguistic representations and semantic gestures are read in by the *Multimodal Dialogue Manager (MDM)*. Some multimodal dialogue managers have been proposed over the years, e.g.: (Wahlster, 2006; Sanders and Holzapfel, 2008; Peternel et al., 2014; Ondáš and Juhár, 2015). In developing the MDM, there are a variety of topics in human-robot communication to address. For a review, see: (Breazeal et al., 2004; Tellex et al., 2011; Ajoudani et al., 2018; Gluck and Laird, 2019).

We take inspiration from the above cited multimodal dialogue managers in combination with a more recently implemented open-source dialogue manager called VOnDa, (Kiefer et al., 2019). Dialogue management using VOnDa is founded on the information state based approach (Traum and Larsson, 2003). The information state contains the robot's state, including dialogue as well as domain specific information. Here, the information state may be extended by additional [multimodal] contextual knowledge. VOnDa's information state is represented as extended OWL ontologies and managed using a semantic repository and reasoner called HFC (Krieger and Willms, 2015). With VOnDa, changes in the robot's information state trigger a declarative rule system with statistical selection to generate a *dialog act* in response to the situation. A dialogue act generally results in

the output of text (to be converted to speech), but may also be realized as motor control directives, and other modalities, such as affective cues for a text-to-speech synthesizer. For the MDM we are also pursuing how to incorporate a construction grammar approach with ontologies for language learning. See: (Steels, 2004; Oliva et al., 2012; Lindes and Laird, 2017). We are also considering how our MDM may integrate with a VoxML approach (Pustejovsky and Krishnaswamy, 2016).

Output from the MDM is combined with output from the TM to assemble a control signal by the *Interaction Planner (IP),* to be interpreted and executed by the IM. This signal is implemented using an extensible markup protocol. The IM runs locally on the robot and is designed to be very fast, mainly handling interrupts and conflict resolution. Meanwhile, the CCA may be hosted on a super machine or distributed across machines with unlimited processing power. Though the control signal from the CCA via the IP is dynamically generated, stand-alone or static control scripts may be used in place of the CCA. This allows the IM and its processing modules to be tested in the absence of the CCA. This also allows the IM to be developed as a stand-alone Robot Operating System (ROS) package, to be used with other cognitive architectures. The use of static control scripts in place of the CCA converts our system architecture into a menu-driven dialogue system. That is, with static control scripts the IM may be regarded as something of a multimodal VoiceXML interpreter.

Consider the following scenario. A human trainer named John begins a learning session by saying something along the lines of "okay robot, let's learn a new gesture." With this, the robot is triggered to enter 'gesture learning mode' and when the robot is ready with its front camera recording, the robot responds with some variation of "okay, John, I'm ready." John then performs the body gesture that he wants the robot to learn. For example, let us consider a gesture to indicate 'stop' - the gesture a police officer might use when directing traffic and signaling a car to stop (as in Figure 1, bottom left). While performing the gesture, John may give a verbal description, such as "lift your hand like this, palm up and fingers stretched, and extend the arm forward." Once John has finished producing the gesture, he then says: "that's it," and the robot acknowledges this by saying "okay," or something analogous. The video recording of the gesture is then saved and processed into a labeled representation as described in Section 2.2.

After processing and maybe after multiple examples of the desired gesture have been recorded, the robot should be ready to produce the gesture. In this case, the robot says something amounting to: "shall I perform the gesture now?" and John may respond with feedback indicating "yes" or "no," prompting the robot to then execute the gesture or not. If there was a problem during processing, the robot may ask John to repeat the gesture. Once the robot has performed the gesture, the robot then asks: "was that okay?" and John may verbally respond "yes, good" while nodding his head 'yes' and-or giving a 'thumbs up' gesture. Or John may indicate 'no, let's try again' while shaking his head 'no' and giving a 'thumbs down' hand gesture (assuming yes/no head and hand gestures have been acquired by the robot). Either a verbal command or a visual command should be enough for the interaction to proceed. The robot might then say 'what does this gesture mean?' John would then explain the meaning of the gesture and the robot would store the gesture with a semantic label (e.g. 'stop').

2.2 Gesture Acquisition

When in 'gesture recording mode,' the robot records a video of the person's complete motion. Each motion or gesture is stored in a buffer as a video example. The trainer (or multiple different trainers) can record the same motion multiple times, and the repetitions are stored as new examples under the same class. We use OpenPose (Cao et al., 2019) for its current superior performance in extracting 2D skeletal information from the recorded video examples. For representing and reconstructing 3D motions from the 2D poses, we deploy a dilated fully convolutional model (Pavllo et al., 2019) to estimate a 3D skeletal pose at each sampled frame. Each pose is represented as a set of Cartesian joint positions. Sequences of the extracted 3D positions are transformed into estimated motor positions for a single video example, and are saved as a *motor trajectory*. A motor trajectory takes the form of a matrix. The columns of the matrix correspond to motor channels of the robot, and rows of the matrix correspond to the passage of time. If a user is satisfied with a gesture reproduced by the robot, the video sample of the gesture may be discarded, and only the motor trajectory needs to be saved.

Though the robot can produce a gesture based on a single example, it is better to generalize the motion under the same gesture label, assuming there are multiple examples for the same class of gesture. This is done to reflect naturalness that real people perform the same motion with a rich repertoire of variations. In order to capture these variations, we apply a mixture of Gaussians (Min and Chai, 2012) to generalize the distribution of the motion examples $P(\mathbf{x})$ for each gesture. This is done following Equation 1.

$$P(\mathbf{x}) = \sum_{k=1}^{K} \phi_k N(\mu_k(x), \sigma_k(x)) \quad (1)$$

One issue in combining multiple motor trajectories is that each motion example may have a different length, meaning the number of frames could vary. To address this, we define a canonical timeline and time normalize all motion examples in the same class to this canonical timeline. The resulting *statistical motion model* provides a compact way to represent each gesture as a set of discrete examples. With statistical motion models, gestures can be represented in a continuous manifold space. In the gesture production phase, if the robot is asked to perform a gesture (e,g. 'stop') without any additional constraints, our model can sample a random motion to be close to the examples with high likelihood. For the gestures with additional constraints, for instance, if the direction of the robot arm is specified, or the robot starts from an unusual initial pose, our model can formulize it as an optimization problem to find the best match in a continuous motion space. Following Equation 2.

$$\arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{c}) \quad (2)$$

where $\mathbf{c}$ is a set of constraints, which can be target positions or orientations, and even some high level constraints. Furthermore, if an end effector position is specified, the statistical motion model can be coupled with inverse kinematics and-or a visual guidance system. Our system does not simply produce deterministic motions from examples, but is enabled to produce similar motions with new variations. In addition, our motion model can be continuously tuned by adding new examples.

It should be noted that in estimating motor positions from Cartesian 3D joint data using inverse kinematics, there is 'motor bleed over.' This relates to how people's skeletons differ in size and proportion to each other and to the robot's skeleton. It is thus difficult to isolate desired robot motor movements for system calibration. An improved method for motor position estimation from skeletal data is desired and is a focus of current efforts.

3 Discussion

We have introduced the infrastructure of an interactive speech-vision-motor system for training a life-sized humanoid robot to produce desired arm and head gestures. The system interfaces a rudimentary cognitive architecture with an interaction manager for robot control. We use an LfD technique combined with spoken instructions and dialogue for training a robot to produce gestures. We lastly turn to consider the relationship between perception and action, the language grounding problem, and semantic representation.

There is an intimate relationship between perception and action. The research industry surrounding the mirror neuron hypothesis reifies this (Hickok, 2014) In light of this, our current work also includes the development of a gesture recognition algorithm that depends on production learning. The time-normalized motor trajectories of a class from Section 2.2 define a centroid motor trajectory for the class. We call this centroid a *gesture prototype*. In short, a motor trajectory to be categorized is template-matched against the stored inventory of gesture prototypes using a multidimensional dynamic time warping algorithm (Müller, 2007). The best match is taken as the gesture's category.

Plans are to develop our system to addresses the symbol grounding problem (Harnad, 1990; Steels, 2003; Cangelosi, 2010; Misra et al., 2016). Establishing a socially situated and embodied system for interactive gesture learning was but a first step. Semantic meaning must be grounded in experience, where different modalities (speech, vision, motor feedback) are integrated. Interactive audio-visual-motor recordings from our system may be used for machine learning approaches, e.g. (Santín et al., 2020) to train multi-modal speech recognizers. In order for meaning to emerge, the robot must 'understand' its own output. By pursuing a paradigm where gesture recognition is based on the robot's representations for gesture production, our hope is to depict representations to be one and the same for perception and production. In viewing speech as a problem of motor control, speech cognition becomes grounded in the robot's experience.

Acknowledgments

We would like to thank our colleagues for discussions and three anonymous reviewers for their feedback. This work is part of the research project XAINES, funded by grant No. 01IW20005 of the German Ministry for Education and Research (BMBF). Please find more information on XAINES here: https://www.dfki.de/en/web/research/projects-and-publications/projects-overview/projekt/xaines

References

Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. 2018. Progress and prospects of the human–robot collaboration. *Autonomous Robots*, 42(5):957–975.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483.

Michael Connolly Brady. 2016. A low cost desktop robot and tele-presence device for interactive speech research. In *INTERSPEECH, 2016, San Francisco*.

Cynthia Breazeal, Guy Hoffman, and Andrea Lockerd. 2004. Teaching and working with robots as a collaboration. In *AAMAS*, volume 4, pages 1030–1037.

Sylvain Calinon and Aude Billard. 2007. Learning of gestures by imitation in a humanoid robot. Technical report, Cambridge University Press.

Sylvain Calinon, Florent D'halluin, Eric L Sauser, Darwin G Caldwell, and Aude G Billard. 2010. Learning and reproduction of gestures by imitation. *IEEE Robotics & Automation Magazine*, 17(2):44–54.

Angelo Cangelosi. 2010. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of life reviews*, 7(2):139–151.

Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. 2010. Robust spoken instruction understanding for hri. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 275–282. IEEE.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.

Laurence Devillers, Tatsuya Kawahara, Roger K Moore, and Matthias Scheutz. 2020. Spoken language interaction with virtual agents and robots (slivar): Towards effective and ethical interaction (dagstuhl seminar 2021). In *Dagstuhl Reports*, volume 10. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Guglielmo Gemignani, Emanuele Bastianelli, and Daniele Nardi. 2015. Teaching robots parametrized executable plans through spoken interaction. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 851–859.

Kevin A Gluck and John E Laird. 2019. Interactive task learning. *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*, 26:1.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Gregory Hickok. 2014. *The myth of mirror neurons: The real neuroscience of communication and cognition*. WW Norton & Company.

Bernd Kiefer, Anna Welker, and Christophe Biwer. 2019. Vonda: A framework for ontology-based dialogue management. *In International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Nathan Koenig, Leila Takayama, and Maja Matarić. 2010. Communication and knowledge sharing in human–robot interaction and learning from demonstration. *Neural Networks*, 23(8-9):1104–1112.

Hans-Ulrich Krieger and Christian Willms. 2015. Extending owl ontologies by cartesian types to represent n-ary relations in natural language. In *Proceedings of the 1st Workshop on Language and Ontologies*.

Ivana Kruijff-Korbayova, Johannes Hackbarth, Caspar Jacob, Bernd Kiefer, Matthias Schmitt, Tanja Schneeberger, Tim Schwartz, Hanns-Peter Horn, and Karsten Bohlmann. 2020. Towards intuitive verbal and non-verbal communication for incidental robot-human encounters in clinic hallways. In *Astrid Rosenthal-von der Ptten, David Sirkin, Anna Abrams, Laura Platte (editor). Workshop on Incidental encounters with Robots in Public Spaces, Cambridge United Kingdom Aachen University*.

Jangwon Lee. 2017. A survey of robot learning from demonstrations for human-robot collaboration. *arXiv preprint arXiv:1710.08789*.

Peter Lindes and John E Laird. 2017. Cognitive modeling approaches to language comprehension using construction grammar. In *2017 AAAI Spring Symposium Series*.

Rui Liu and Xiaoli Zhang. 2019. A review of methodologies for natural-language-facilitated human–robot cooperation. *International Journal of Advanced Robotic Systems*, 16(3):1729881419851402.

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

Jianyuan Min and Jinxiang Chai. 2012. Motion graphs++ a compact generative model for semantic motion analysis and synthesis. *ACM Transactions on Graphics (TOG)*, 31(6):1–12.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3d environments with visual goal prediction. *arXiv preprint arXiv:1809.00786*.

Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. 2016. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300.

Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Jesús Oliva, Jerome Feldman, Luca Gilardi, and Ellen Dodge. 2012. Ontology driven contextual best fit in embodied construction grammar. In *International Workshop on Constraint Solving and Language Processing*, pages 133–151. Springer.

Stanislav Ondáš and Jozef Juhár. 2015. Event-based dialogue manager for multimodal systems. In *Emergent Trends in Robotics and Intelligent Systems*, pages 227–235. Springer.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762.

Luka Peternel, Tadej Petrič, Erhan Oztop, and Jan Babič. 2014. Teaching robots to cooperate with humans in dynamic manipulation tasks based on multi-modal human-in-the-loop approach. *Autonomous robots*, 36(1):123–136.

James Pustejovsky and Nikhil Krishnaswamy. 2016. Voxml: A visualization modeling language. *arXiv preprint arXiv:1610.01508*.

Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. 2020. Recent advances in robot learning from demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:297–330.

David Sanders and Hartwig Holzapfel. 2008. A dialogue manager for multimodal human-robot interaction and learning of a humanoid robot. *Industrial Robot: An International Journal*.

José Miguel Cano Santín, Simon Dobnik, and Mehdi Ghanimifard. 2020. Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 53–61.

Lanbo She, Yu Cheng, Joyce Y Chai, Yunyi Jia, Shaohua Yang, and Ning Xi. 2014. Teaching robots new actions through natural language instructions. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 868–873. IEEE.

Luc Steels. 2003. Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312.

Luc Steels. 2004. Constructivist development of grounded construction grammars. ACL '04. Association for Computational Linguistics.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer.

Wolfgang Wahlster. 2006. Dialogue systems go multimodal: The smartkom experience. In *SmartKom: foundations of multimodal dialogue systems*, pages 3–27. Springer.

Tom Williams and Matthias Scheutz. 2017. The state-of-the-art in autonomous wheelchairs controlled through natural language: A survey. *Robotics and Autonomous Systems*, 96:171–183.

Zuyuan Zhu and Huosheng Hu. 2018. Robot learning from demonstration in robotic assembly: A survey. *Robotics*, 7(2):17.