

Multimodal Interactive Learning of Primitive Actions

Tuan Do, Nikhil Krishnaswamy, Kyeongmin Rim, and James Pustejovsky

Department of Computer Science

Brandeis University

Waltham, MA 02453 USA

{tuandn, nkrishna, krim, jamesp}@brandeis.edu

Abstract

We describe an ongoing project in learning to perform primitive actions from demonstrations using an interactive interface. In our previous work, we have used demonstrations captured from humans performing actions as training samples for a neural network-based trajectory model of actions to be performed by a computational agent in novel setups. We found that our original framework had some limitations that we hope to overcome by incorporating communication between the human and the computational agent, using the interaction between them to fine-tune the model learned by the machine. We propose a framework that uses multimodal human-computer interaction to teach action concepts to machines, making use of both live demonstration and communication through natural language, as two distinct teaching modalities, while requiring few training samples.

Introduction

This work takes a position on learning primitive actions or interpretations of low-level motion predicates by the Learning from Demonstration (LfD) approach. LfD can be traced back to the 1980s, in the form of automatic robot programming (e.g., Lozano-Perez (1983)). Early LfD is typically referred to as *teaching by showing or guiding*, in which a robot’s effectors could be moved to desired positions, and the robotic controller records its coordinates and rotations for later re-enactment. In this study, we instead focus on a methodology to teach action concepts to computational agents, allowing us to experiment with a proxy for the robot without concern for physically controlling the effectors.

As discussed in Chernova and Thomaz (2014), there are typically two sub-categories of actions that can be taught to robots: 1) high-level tasks that are hierarchical combinations of lower-level motion trajectories; and 2) low-level motion trajectories, the focus of this study, that can be taught by using a feature-matching method. We have experimented with offline learning motion trajectories from captured demonstrations. This method has some limitations, including requiring multiple samples as opposed to one-shot (or few-shot) learning, and being unable to accept corrections to generated examples beyond training on more data (Do, Krishnaswamy, and Pustejovsky, 2017).

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

There is a wealth of prior research on hierarchical learning of complex tasks from simpler actions (Veeraraghavan, Papanikolopoulos, and Schrater, 2007; Dubba et al., 2015; Wu et al., 2015; Alayrac et al., 2016; Fernando, Shirazi, and Gould, 2017). Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) have been used extensively in previous work (Akgun et al., 2012; Calinon and Billard, 2007) to model the learning and the reenacting components. Do (2018) proposed to use Reinforcement Learning (RL) directed by a *shape* rewarding function learned from sample trajectories. In contrast, we are investigating LfD methods to teach primitive concepts such as *move A around B*, or *lean A against B*, or *build a row*, or *build a stack from blocks on the table*, and we propose a method to learn these action concepts from demonstrations, supplemented by interaction with the agent to verify or correct some of the suppositions that the agent learns while building a demonstration-trained model.

Recently, Mohseni-Kabir et al. (2018) proposed a methodology to jointly learn primitive actions and high-level tasks from visual demonstrations, with the support of an interactive question-answering interface. In this framework, robots ask questions in order to group primitive actions together to create high-level actions. In a similar fashion, Lindes et al. (2017) teach a task to robots, such as *discard an object*, by giving step-by-step instructions built on top of the simple actions *move*, *pick up*, *put down*; Maeda et al. (2017) showcase a system wherein a robot makes active requests and decisions in the course of learning primitive actions incrementally; Tellex et al. (2011) use probabilistic graphical models to ground natural language commands to the situation. We think these types of communicative frameworks can be extended to learning low-level actions.

We view this direction of interactive learning as particularly promising, where symmetric communication between humans and robots can be used to complement LfD as a modality for teaching (cf. Thomaz and Breazeal (2008)).

Related research

Naturalistic communication between humans tends to be multimodal (Veinott et al., 1999; Narayana et al., 2018). Human speech is often supplemented by non-verbal communication (gestures, body language, demonstration/“acting”), while linguistic expressions provide both transparent and

abstract information regarding the actions and events in the situation, much of which is not readily available from demonstrations. Dynamic event structure (Pustejovsky and Moszkowicz, 2011; Pustejovsky, 2013) is one approach to language meaning that formally encodes events as programs in a dynamic logic with an operational semantics. These events very naturally map to the sub-steps undertaken during the course of demonstrating a new action (“grasp”, “pick up”, “move to location”, etc.). Motion verbs can be divided into complementary *manner-* or *path-*oriented predicates and adjuncts (Jackendoff, 1983). Changes over time can be neatly encapsulated in durative verbs as well as in gestures or deictic referents denoting trajectory and direction. This allows humans to express where an object should be or go either using linguistic descriptions or by directly indicating approximate paths and locations.

Computational agents typically lack the infrastructure required to learn new concepts solely through linguistic description, often due to an inability to fully capture the intricate semantics of natural language. Thus, instead of providing verbose instructions, we treat the agent as an active learner who interacts with the teacher to understand new concepts, as suggested in Chernova and Thomaz (2014).

Research in cognition (cf. Agam and Sekuler (2008)) has investigated how humans imitate trajectories of different shapes, giving a strong indication that we tend to be able to better remember trajectories that follow a consistent pattern (*curvature consistency*). In this paper, we hypothesize that human primitive action concepts exhibit relatively transparent conceptual consistencies. We hope to learn these consistencies directly from data represented as sequential features on a frame-by-frame basis from demonstrations.

Qualitative spatial (QS) representations have proven useful in analogical reasoning, allowing machine learning algorithms to perform generalizations over smaller amounts of data than required for traditional quantitative representations (McLure, Friedman, and Forbus, 2015). This allows it to serve as a bias in the model, reflecting human real-world knowledge as the human interlocutor would be expected to have. Libraries of qualitative relations often draw extensively from longstanding observations of human bias in psychological experiments on spatial processing (Stevens and Coupe, 1978; Gentner and Collins, 1981; Weld and De Kleer, 2013). Thus a machine learning model trained on data that contains bias from humans should reflect those same biases and the model should demonstrate judgments of those of human subjects, in spatial reasoning as in other domains (Caliskan, Bryson, and Narayanan, 2017). In a limited sense, biases can be defined as modes (in a statistical sense) of a non-uniform distribution in a descriptive space. For example, our bias toward the value of 0 (in a range from 0 to ∞) allows us to qualitatively distinguish between when two objects are *externally connected* (i.e., with 0 or near-0 distance between them) or *disconnected*.

To summarize, we exploit the ability to express and describe actions in multiple modalities, in order to explore how to impart human concepts and biases to AI models. We use the QS representations as the abstraction level to bridge demonstrative data, action concepts, and the action language

used and expressed by computational agents.

A learning scenario

In the following conversation, both a human user and an avatar can observe a table containing both a red block and a green block:

HUMAN: Can you **move the red block around the green block**?

ROBOT: Can you show me what “around” means?

HUMAN: Yes. [*Slides the red block clockwise around the green block on the table*]

ROBOT: Is the green block always stationary?

HUMAN: Yes.

ROBOT: Is the red block always moving?

HUMAN: Yes.

ROBOT: Is the red block always about the same distance from the green block?

HUMAN: Yes.

ROBOT: Does the red block always move in the same direction relative to the green block?

HUMAN: Yes.

ROBOT: OK. Let me slide the red block around the green block.

[*Performs the action **slide the red block around the green block**, clockwise*]

This scenario displays an effort to teach a machine a concept that is rather difficult to learn given only one or a few demonstrative samples. We also want to demonstrate the *desiderata* of a machine learning system that can facilitate that learning:

- It can recognize pattern consistencies from feature data. Consistencies should be in formulaic representations that can be clearly articulated in natural language expressions.
- Pattern consistencies can be evaluated over multiple frames of the same demonstration. More importantly, a desirable framework should allow us to estimate the confidence of a pattern intended by the instructor.
- The system should take a proactive role in interaction, by asking questions pertaining to patterns need to be verified.
- In terms of natural language interaction, the system has to be able to identify novel ideas as missing concepts in its semantic framework as well as to generate questions for verification of the recognized patterns.

Framework

Figure 1 depicts the architecture of our learning system. For the top component, our experimental setup makes use of simple markers attached to objects for recognition and tracking. For natural language grounding, our proposal leverage the advancement of speech recognition (Povey et al., 2011) and syntactic analysis tools (Chen and Manning, 2014; Reddy et al., 2017) to generate a grounded interpretation from spoken language. For the bottom component, we discuss the use of “mined patterns” as constraints for action reenactment.

The focus in this section is the middle component (inside the dotted box), including methods to mine pattern consistencies from demonstrative data, then pose generated natural

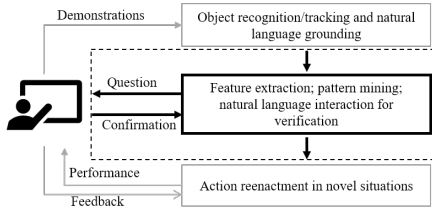


Figure 1: Interactive learning framework

language questions to human teachers to ask for confirmation of conceptual understanding, and use this understanding to constrain action performance when presented with a novel context or setup.

Representations

To represent pattern consistencies, we use a set of qualitative features that are widely used in the Qualitative Spatial Reasoning (QSR) community. We have used these features as representation for action recognition (Do and Pustejovsky, 2017) This is not intended to be an exhaustive set of features, and other feature sets, such as the Region Connection Calculus (RCC) (Cohn et al., 1997), could be used as well.

- **CARDINAL DIRECTION (CD)** (Andrew, Mark, and White, 1991), transforms compass relations between two objects into canonical directions such as North, North-east, etc., producing 9 different values, including one for where two locations are identical. This feature can be used for the relative direction between two objects, for an object’s orientation, or for its direction of movement.
- **MOVING or STATIC (MV)** measures whether a point is moving or not.
- **QUALITATIVE DISTANCE CALCULUS (QDC)** discretizes the distance between two moving points, e.g., the distance between two centers of two blocks.
- **QUALITATIVE TRAJECTORY CALCULUS (Double Cross): QTC_C** is a representation of motions between two objects by considering them as two moving point objects (MPOs) (Delafontaine, Cohn, and Van de Weghe, 2011). We consider two feature types of this set, whether two points are moving toward each other (QTC_{C1}) or whether they are moving clockwise or counterclockwise w.r.t. each other (QTC_{C3}).

These qualitative features can be used to create the formulaic pattern consistencies that we are looking for in the previous discussion. All features can be interpreted as univariate or multivariate functions binding to tracked objects at a certain time frame. Hereafter, let $f_t^k(d)(x, y)$ denote the qualitative feature extracted from demonstration d at frame t of feature type k between two objects x and y .

Pattern mining

The following describes some of the pattern consistencies that we are hoping to learn from data:

- $f_0^k(d)(x, y) ? \alpha$ where $?$ can be any comparison operator $<, >, =, \leq, \geq, \neq$, and α is a constant value. This is a state to be satisfied at the start of a demonstration.
- $f_F^k(d)(x, y) ? \alpha$ is a *final* (F) state to be satisfied at the end of a demonstration.

- $\forall t f_t^k(d)(x, y) ? \alpha$ describes a feature value that stays constant across all frames.
- $\forall t f_t^k(d)(x, y) ? f_{t+1}^k(d)(x', y')$ describes a feature relationship between two consecutive frames. We allow a form of dynamic object binding so that it is not necessary that $(x, y) = (x', y')$, i.e., object binding is made by evaluating the demonstration d at time t . However, in the example of “Slide A around B”, (x, y) always bind to (A, B), because the system can map these directly from the instruction given to the demonstration.
- $f_0^k(d)(x, y) ? f_F^k(d)(x', y')$ relates features at the start (frame 0) and end (frame F) of the demonstration.

These patterns $p \in \mathcal{P}$, where \mathcal{P} is a partially ordered set. We define a *precedence* relation, \preceq , so that two patterns can be compared. $p_1 \preceq p_2$ if p_1 is logically superseded by p_2 . For example, $p_1 = f_0^k(d)(x, y) < \alpha$ takes precedence over $p_2 = f_0^k(d)(x, y) \neq \alpha$.

To detect these patterns from data, we can define a function over patterns $q(p)$ that measures how confident we are that a pattern is intended in an action concept. This value should be higher when we have more demonstrations that exhibit the same pattern. Furthermore, q should also give a pattern with a higher precedence a higher salience. The intuition is that if $p_1 \preceq p_2$, and we have $q(p_1) > t \wedge q(p_2) > t$ where t is a confidence threshold, the system should ask for confirmation about p_1 before asking about p_2 . When the teacher confirms p_1 to be true, the system then can take p_2 as trivially true.

Though attaining such a function is not trivial, we will give an illustrative example. Assume a 4-part quantization of QDC (“adjacent”, “close”, “far”, “very far”), we define a bias b over these values that characterizes the likelihood of a quantized region v to be recognizable, for example $b = 1/v$. Finally, let $domain(p)$ be the range of the feature function f that p uses. Now, we define a heuristic function $q(p)$ as follows:

$$q(p) = \frac{probability(p) * bias(domain(p))}{|domain(p)|}$$

whereas $probability(p)$ is the probability that p is correct among all samples, $bias(domain(p)) = \sum_v b(v)$, and $|domain(p)|$ is the size of the domain. For example, if in 80% of the samples, $f = 0$ and in the remaining 20%, $f = 1$, we have $q(f = 0) = 0.8 * 1/1 = 0.8$, $q(f \leq 1) = 1 * (1 + 0.5)/2 = 0.75$, therefore, $q(f \leq 1) < q(f = 0)$. If the ratio is 50:50, $q(f \leq 1) > q(f = 0)$.

Generating natural language questions

Now that the system has patterns of qualitative features from observations, each associated with a *confidence* score of consistency from the function q , it needs to confirm the intentionality of the patterns with the teacher. To come up with a proper set of questions to ask, first the patterns need to be arranged into a queue in an order using the *precedence* relation and confidence value, then the system forms natural language questions from the queue of patterns that need to be confirmed. When the system gets a confirmation, it will iterate through the queue to remove now-implicit patterns.

For instance, suppose p_1 “object X moving to the east” and p_2 “ X moving in one direction” where $p_1 \preceq p_2$. When the teacher confirms p_1 first, the system does not have to ask for confirmation on p_2 . However, if the system is given only one demonstration, the function q might not assign high confidence to $p_1 \equiv \forall t f_t^{MV.dir}(d)(X) = EAST$ but higher value to $p_2 \equiv \forall t (f_t^{MV.dir}(d)(X) = f_{t+1}^{MV.dir}(d)(X))$ based on specificity level of p_1 ; p_2 does not rely on actual values to which the feature evaluates. In such a case the system prioritizes p_2 in the confirmation queue, if p_1 gets enqueued at all with confidence threshold.

For linguistic translation, we use a mapping from qualitative features and time interval of patterns to linguistic descriptions that, in turn, are composed to complete natural language sentences using the rule-based slot-filling mechanism in the interactive interface (Krishnaswamy and Pustejovsky, 2016) (see below).

Performing actions in novel situations

In previous work (Do, 2018), we have addressed a few different approaches to performing learned actions in a novel situations, e.g., RL and search algorithms. We must incorporate the learned pattern consistencies from the previous discussion as constraints in the search space of the execution planning. The algorithm we use for action reenactment is a search algorithm, in which an execution is a chain of planned simple steps, such as “Move(A , $coordinate$)” or “Rotate(A , $angle$)”. A simple search algorithm will generate a set of random candidate steps on the search space, then qualify whether a new step satisfies the constraints. At the same time, the system can verify if a state satisfies a termination condition and decide to announce the completion of the action. *Best-first* or *beam* searches can both be used in this scenario.

Experiments



Figure 2: Sample interaction with avatar

Our interaction framework is built on the VoxML/VoxSim platform (Pustejovsky and Krishnaswamy, 2016; Krishnaswamy and Pustejovsky, 2016), which facilitates the encoding of common-sense knowledge about events and objects, and their visualization in a 3D environment as a demonstration of how the computer interprets them. The interaction system is presented in full by Krishnaswamy et al. (2017). A human user, standing before a Kinect® camera and a monitor that displays an avatar and a table with blocks on it can, through the use of language and gesture, direct the avatar to perform a set of actions with the blocks. For instance, by indicating a block (e.g., by saying “the purple block” or pointing to the purple block), and then pointing to a new location, the user can direct the avatar to move that block to that new location.

At points in the interaction, the avatar may express uncertainty about an action or need to confirm it. This may

be asking which block the user is indicating, or confirming that the indicated location is the intended destination of the block. When the avatar encounters a symbol sequence such as $point_at(block7); point_at(L1)$ with the existing instruction that the block is to be slid, it can insert these symbols into a predicate structure, e.g., $slide(block7, L1)$ and then disambiguate any piece that it needs to by translating those symbols into a natural language output. On this primitive level, these simple symbols are encoded in the model, and so with a learning-based approach, the question becomes one of being able to extract more complex symbols in need of confirmation or disambiguation from a learning algorithm, determining what to ask about, and how to phrase the question to prompt a “yes” or “no” answer.

Each time the avatar completes an action (i.e., finishes moving an object), we write out the complete state of the scene, with the positions and rotations of all blocks. From this raw data, we extract qualitative spatial relations to be fed into the learning module, using QSRLib (Gatsoulis et al., 2016).

Some examples of patterns we expect to mine from action demonstrations are given here:

Move A around B

- $\forall t f_t^{MV}(A) = 1$ and $\forall t f_t^{MV}(B) = 0$
- $\forall t f_t^{QTCc3}(A, B) = f_{t+1}^{QTCc3}(A, B)$
- $f_0^{CD}(A, B) = f_F^{CD}(A, B)$

Make a row of blocks: we assume that a “row of blocks” means blocks that are evenly spaced along a single axis. Let us assume that all blocks make a set S . Let us define some functions on S (that we called dynamic binding in the previous section): $\mathcal{L}(S)$ is the last moved block in S ; $\mathcal{C}(S)$ is a selected block for the next move; then we have the follows:

- $\forall t f_t^{QDC}(\mathcal{L}(S), \mathcal{C}(S)) = f_{t+1}^{QDC}(\mathcal{L}(S), \mathcal{C}(S))$
- $\forall t f_t^{CD}(\mathcal{L}(S), \mathcal{C}(S)) = f_{t+1}^{CD}(\mathcal{L}(S), \mathcal{C}(S))$

Conclusion and Discussion

Currently, we are working with data captured from the interactive interface, in which a demonstration has already been broken down into multiple steps (Figure 2). To extend this framework to work with real demonstrations, where data comes in from a continuous stream, we can capture data from real human performances using tools such as ECAT (Do, Krishnaswamy, and Pustejovsky, 2016). We may require processing to facilitate pattern mining, including noise removal or “key” frame detection (cf. Asfour et al. (2008)).

It is also important that the system be able to recognize the relationship between different action concepts. Taking two action concepts, (1) “Move A around B” and (2) “Move A around B clockwise”, as examples, we can see the hierarchical relationship between them implies their corresponding conceptual patterns. Therefore, our system needs to be able to update the learned concept of the first action to be a superclass of the concept of the second action.

We have proposed a system for learning primitive actions using an interactive learning interface. By examining a specific learning scenario, we have demonstrated various requirements of this system. We believe such a framework can improve on the successes of LfD methods by incorporating multimodal information through real-time interaction to facilitate online learning from sparse data to improve models.

References

- Agam, Y., and Sekuler, R. 2008. Geometric structure and chunking in reproduction of motion sequences. *Journal of Vision* 8(1):11–11.
- Akgun, B.; Cakmak, M.; Yoo, J. W.; and Thomaz, A. L. 2012. Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 391–398. ACM.
- Alayrac, J.-B.; Bojanowski, P.; Agrawal, N.; Sivic, J.; Laptev, I.; and Lacoste-Julien, S. 2016. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4575–4583.
- Andrew, U.; Mark, D.; and White, D. 1991. Qualitative spatial reasoning about cardinal directions. In *Proc. of the 7th Austrian Conf. on Artificial Intelligence. Baltimore: Morgan Kaufmann*, 157–167.
- Asfour, T.; Azad, P.; Gyarfas, F.; and Dillmann, R. 2008. Imitation learning of dual-arm manipulation tasks in humanoid robots. *International Journal of Humanoid Robotics* 5(02):183–202.
- Calinon, S., and Billard, A. G. 2007. What is the teachers role in robot programming by demonstration?: Toward benchmarks for improved learning. *Interaction Studies* 8(3):441–464.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.
- Chen, D., and Manning, C. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 740–750.
- Chernova, S., and Thomaz, A. L. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8(3):1–121.
- Cohn, A. G.; Bennett, B.; Gooday, J.; and Gotts, N. M. 1997. Representing and reasoning with qualitative spatial relations about regions. In *Spatial and temporal reasoning*. Springer. 97–134.
- Delafontaine, M.; Cohn, A. G.; and Van de Weghe, N. 2011. Implementing a qualitative calculus to analyse moving point objects. *Expert Systems with Applications* 38(5):5187–5196.
- Do, T., and Pustejovsky, J. 2017. Learning event representation: As sparse as possible, but not sparser. *IJCAI Qualitative Reasoning Workshop*.
- Do, T.; Krishnaswamy, N.; and Pustejovsky, J. 2016. ECAT: Event capture annotation tool. *Workshop on Interoperable Semantic Annotation (ISA)*.
- Do, T.; Krishnaswamy, N.; and Pustejovsky, J. 2017. Teaching virtual agents to perform complex spatial-temporal activities. *AAAI Spring Symposium: Integrating Representation, Reasoning, Learning, and Execution for Goal Directed Autonomy*.
- Do, T. 2018. *Learning to perform actions from demonstrations with sequential modeling*. Ph.D. Dissertation, Brandeis University.
- Dubba, K. S.; Cohn, A. G.; Hogg, D. C.; Bhatt, M.; and Dylla, F. 2015. Learning relational event models from video. *Journal of Artificial Intelligence Research* 53:41–90.
- Fernando, B.; Shirazi, S.; and Gould, S. 2017. Unsupervised human action detection by action matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–9.
- Gatsoulis, Y.; Alomari, M.; Burbridge, C.; Dondrup, C.; Duckworth, P.; Lightbody, P.; Hanheide, M.; Hawes, N.; and Cohn, A. 2016. Qsrlib: a software library for online acquisition of qualitative spatial relations from video. In *Workshop on Qualitative Reasoning (QR16), at IJCAI-16*.
- Gentner, D., and Collins, A. 1981. Studies of inference from lack of knowledge. *Memory & Cognition* 9(4):434–443.
- Jackendoff, R. 1983. *Semantics and Cognition*. MIT Press.
- Krishnaswamy, N., and Pustejovsky, J. 2016. VoxSim: A visual platform for modeling motion language. In *Proceedings the 26th International Conference on Computational Linguistics: System Demonstrations*.
- Krishnaswamy, N.; Narayana, P.; Wang, I.; Rim, K.; Bangar, R.; Patil, D.; Mulay, G.; Beveridge, R.; Ruiz, J.; Draper, B.; et al. 2017. Communicating and acting: Understanding gesture in simulation semantics. In *12th International Conference on Computational Semantics (IWCS), Short papers*.
- Lindes, P.; Mininger, A.; Kirk, J. R.; and Laird, J. E. 2017. Grounding language for interactive task learning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, 1–9.
- Lozano-Perez, T. 1983. Robot programming. *Proceedings of the IEEE* 71(7):821–841.
- Maeda, G.; Ewerton, M.; Osa, T.; Busch, B.; and Peters, J. 2017. Active incremental learning of robot movement primitives. In *Conference on Robot Learning (CORL)*.
- McLure, M. D.; Friedman, S. E.; and Forbus, K. D. 2015. Extending analogical generalization with near-misses. In *AAAI*, 565–571.
- Mohseni-Kabir, A.; Li, C.; Wu, V.; Miller, D.; Hylak, B.; Chernova, S.; Berenson, D.; Sidner, C.; and Rich, C. 2018. Simultaneous learning of hierarchy and primitives for complex robot tasks. *Autonomous Robots*.
- Narayana, P.; Krishnaswamy, N.; Wang, I.; Bangar, R.; Patil, D.; Mulay, G.; Rim, K.; Beveridge, R.; Ruiz, J.; Pustejovsky, J.; and Draper, B. 2018. Cooperating with avatars through gesture, language and action. In *Intelligent Systems Conference (IntelliSys)*.
- Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; Silovsky, J.; Stemmer, G.; and Vesely, K. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.
- Pustejovsky, J., and Krishnaswamy, N. 2016. VoxML: A visual object modeling language. *Proceedings of LREC*.
- Pustejovsky, J., and Moszkowicz, J. 2011. The qualitative spatial dynamics of motion. *The Journal of Spatial Cognition and Computation*.
- Pustejovsky, J. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, 1–10. ACL.
- Reddy, S.; Täckström, O.; Petrov, S.; Steedman, M.; and Lapata, M. 2017. Universal semantic parsing. *arXiv preprint arXiv:1702.03196*.
- Stevens, A., and Coupe, P. 1978. Distortions in judged spatial relations. *Cognitive psychology* 10(4):422–437.
- Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S. J.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, volume 1, 2.

- Thomaz, A. L., and Breazeal, C. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172(6-7):716–737.
- Veeraraghavan, H.; Papanikolopoulos, N.; and Schrater, P. 2007. Learning dynamic event descriptions in image sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–6. IEEE.
- Veinott, E. S.; Olson, J.; Olson, G. M.; and Fu, X. 1999. Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 302–309. ACM.
- Weld, D. S., and De Kleer, J. 2013. *Readings in qualitative reasoning about physical systems*. Morgan Kaufmann.
- Wu, C.; Zhang, J.; Savarese, S.; and Saxena, A. 2015. Watch-n-patch: Unsupervised understanding of actions and relations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4362–4370. IEEE.