COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./DUE DATE **NSF 19-536** **02/19/19**	☐ Special Exception to Deadline Date Policy	**FOR NSF USE ONLY** **NSF PROPOSAL NUMBER** **1925118**

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

IIS - National Robotics Initiative

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
02/19/2019	**1**	**05020000 IIS**	**8013**	**052184116**	**03/15/2019 2:42pm S**

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

250969449

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

Carnegie-Mellon University

AWARDEE ORGANIZATION CODE (IF KNOWN)

0001057000

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

5000 Forbes Avenue
WQED Building
PITTSBURGH, PA 15213-3815

NAME OF PRIMARY PLACE OF PERF

Carnegie-Mellon University

ADDRESS OF PRIMARY PLACE OF PERF, INCLUDING 9 DIGIT ZIP CODE

Carnegie-Mellon University

Pittsburgh ,PA ,152133815 ,US.

IS AWARDEE ORGANIZATION (Check All That Apply) ☐ SMALL BUSINESS ☐ FOR-PROFIT ORGANIZATION ☐ MINORITY BUSINESS ☐ WOMAN-OWNED BUSINESS ☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **NRI: INT: Never-ending Multimodal Collaborative Learning**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **1,499,881**	**48** months	**10/01/19**	

THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☐ BEGINNING INVESTIGATOR
☐ DISCLOSURE OF LOBBYING ACTIVITIES
☐ PROPRIETARY & PRIVILEGED INFORMATION
☐ HISTORIC PLACES
☐ VERTEBRATE ANIMALS IACUC App. Date ______
PHS Animal Welfare Assurance Number ______
☒ TYPE OF PROPOSAL **Research**

☒ HUMAN SUBJECTS Human Subjects Assurance Number **FWA00004206**
Exemption Subsection ______ or IRB App. Date **Pending**
☒ INTERNATIONAL ACTIVITIES: COUNTRY/COUNTRIES INVOLVED
XX
☒ COLLABORATIVE STATUS
Not a collaborative proposal

PI/PD DEPARTMENT

PI/PD FAX NUMBER

PI/PD POSTAL ADDRESS
5000 Forbes Avenue
WQED Building
PITTSBURGH, PA 152133815
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Email Address
PI/PD NAME **Katerina Fragkiadaki**	**DPhil**	**2013**		**kfragki2@andrew.cmu.edu**
CO-PI/PD **Christopher Atkeson**	**PhD**	**1986**	**412-268-5544**	**cga@cs.cmu.edu**
CO-PI/PD **Tom M Mitchell**	**PhD**	**1979**	**412-268-2611**	**Tom.Mitchell@cs.cmu.edu**
CO-PI/PD **Wenzhen Yuan**	**PhD**	**2018**	**857-998-8097**	**wenzheny@andrew.cmu.edu**
CO-PI/PD				

PROJECT SUMMARY

Overview:

NRI: INT: Never-ending Multimodal Collaborative Learning
PI: Katerina Fragkiadaki, Carnegie Mellon University

The proposed work explores and develops algorithms for collaborative and continual perceptual, model, affordance, policy, and reward learning with the assistance of human teachers that employ natural language descriptions paired with visual or kinesthetic demonstrations to teach robotic agents new skills or help them improve and generalize existing ones. The agents jointly learn to ground natural language and acquire new skills by bootstrapping already acquired skills and natural language comprehension, guided by gesticulations and verbal feedback from teachers.

The proposed state and goal representations, forward models, policies, and natural language grounding are all represented using 3D multimodal feature tensors, created by the perceptual front-end of the proposed system. The system is trained using self-supervision to produce geometrically consistent scene models based on view prediction and multimodal coincidence, as well as using supervised learning based on object/attribute and action labels from teachers' instructions. Through experience and interaction with human teachers the agents learn to transform their sensory streams into kinematically consistent and semantically accurate models, and learn to predict the results of their actions and interaction, as well as search over these predictions to find desirable courses of action.

Keywords: Scalability, Customizability, Lowering Barriers, Learning, Perception, Natural Language

Intellectual Merit:

One product of this work will be a rich knowledge base of forward models, generalized policies, objects/attribute/action and reward detectors, natural language parsers and visuotactile state representations that can be used by any robot. The second product will be algorithms to grow and specialize this knowledge base for new situations, tasks, and robots. The central transformative idea of the proposed research is to integrate sensations into 3D feature maps, where entities, objects and parts bind in time and have consistent object-referenced representations independent of viewpoint, as opposed to appear and disappear based on the motion of the observer or other agents or objects.

The proposed representation adds a new spatial dimension to previous feature-based representations, enabling learning robots to utilize spatial reasoning such as SLAM to improve multimodal deep learning. Agents learn to see, reason about temporal evolution, ground natural language, interpret, match and generalize 3D feature representations through continual collaborative learning, which guides them to attend to, focus on, and abstract important parts of the sensory streams. In this way, robots learn to imagine what is behind occlusions and results of actions and natural language goals. They are able to converse and benefit from teachers' feedback through collaborative building of multimodal reasoning, and learn robot- and agent-independent representations of models, affordances, skills, and goals, that support knowledge transfer and knowledge adaptation across heterogeneous robots.

Broader Impacts:

The proposed research will reduce the cost of programming robots and other technology, such as personal assistants. Non-experts will be able to program and personalize robots similarly to how we program fellow humans and especially children: by communicating in natural language (e.g., "stop fidgeting") and demonstrating visually the desired way to do things (e.g., "open it like this"), as opposed to being programmed by writing code or through millions of positive and negative examples. Robots will be able to acquire new concepts and skills adapting to individual users' needs through interaction with end-users, as opposed to maintaining a fixed set of functionalities predetermined at the factory.

The simplicity and directness of grounded natural language interfaces will help robots better serve older adults and people with disabilities. This is just one example of the proposed technology's potential for social good. This research is tightly coupled to the educational program of the PIs, which currently includes a course on language grounding on vision and control, and another on architectures for never-ending learning, with the goal of teaching students that there is more to AI than learning from a large number of positive and negative examples.

TABLE OF CONTENTS

For font size and page formatting specifications, see PAPPG section II.B.2.

Appendix Items:

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

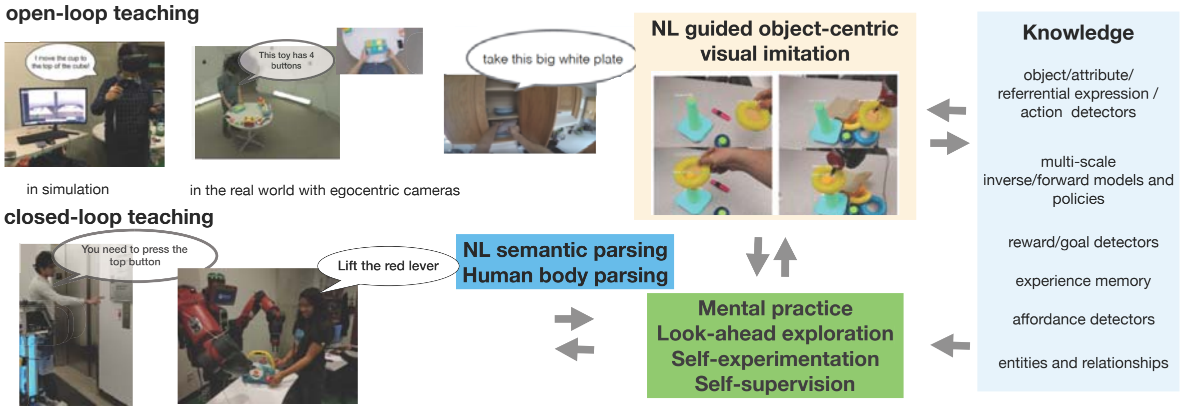

open-loop teaching
I move the cup to the top of the cube!
This toy has 4 buttons
take this big white plate
in simulation
in the real world with egocentric cameras
closed-loop teaching
You need to press the top button
Lift the red lever
NL guided object-centric visual imitation
NL semantic parsing
Human body parsing
Mental practice
Look-ahead exploration
Self-experimentation
Self-supervision
Knowledge
object/attribute/ referential expression / action detectors
multi-scale inverse/forward models and policies
reward/goal detectors
experience memory
affordance detectors
entities and relationships

Figure 1: **Collaborative never-ending multimodal learning:** Humans and robots interact either through live coaching and demonstrations or through offline narrated demonstrations. Self-experimentation and teachers' guidance are distilled into a multitude of diverse object/affordance/attribute/action detectors, reward detectors, goal imaginations and models. Our agents represent their state using 3D feature tensors of visual, tactile and audio features, integrated in time and trained through self-supervision, by predicting results of actions and interactions, or supervised by the language of human teachers. Look-ahead using learned models and affordance detectors accelerate learning and demonstration understanding.

1 Motivation

Humans are social animals and learn from one another through imitation and instruction in a life-long manner. Childhood human learning is a collaborative effort. Children are motivated to learn by innate curiosity [35] and the pleasure of discovering the causes behind events they perceive [36]. Parents and siblings provide a great deal of specialized, shaped, and staged (curricular) input to children and behave in ways that increase stimulation and guidance. Parents and teachers **frequently adjust their children's and students' behaviors** by asking them to say please, lower their voice, sit properly, etc., while at the same time **gesticulating** to help children understand what they mean [77]. In contrast to how humans learn, existing paradigms for programming or learning robotic behaviors are typically non-interactive: updating or extending the robot's learned policies or cost functions during deployment might even require the robot to be withdrawn from operation. **Training often happens once in the lifetime of our robotic agents**, and robotic behaviors cannot be easily customized/personalized to the preferences of a particular user, particular context, or particular environment.

Our proposal puts forward a plan for large-scale **collaborative learning** of affordances, world models, behavioral policies, visuo-tactile representations, and natural language parsing and grounding on heterogeneous robots that are deployed in human environments and continuously learn to adjust and enrich their skill libraries and models (simulators) of the world guided by their internal curiosity and teacher's guidance. Robots can share knowledge much more effectively than humans, which is a major difference between human and robot learning we will exploit. We will address questions of how to most effectively share knowledge across heterogeneous robots from a wide variety of human teachers and learning experiences, including both simulated and actual robot behavior. We will enlist human teachers through fun videogame-like web and virtual reality interfaces, and utilize teachers present in the robot's workspace such as in a hallway. Robots will also be able to seek out clarification or more instruction by requesting web-based or in-person teaching, and communicate with the teacher by replaying what happened. The teachers will use paired natural language (NL), visual input, and, optionally, kinesthetic input (touching and guiding the robot's end-effectors in virtual or actual reality), to demonstrate affordances of objects and tools as well as skills. We call these **narrated demonstrations** (NDs): visual demonstrations synchronized with natural language descriptions.

We address the emphasis and thrusts of the NRI program in the following ways: We address **scalability** by exploring how several heterogeneous robots can learn and pool knowledge from many human teachers. We address **customizability** by enabling robots to learn about the preferences of individual humans as well as aggregating knowledge from many teachers. We address **lowering barriers to entry** by creating low-cost open-source robot designs and engaging tasks for those robots, that are easy for schools and museums to replicate, use, adapt, and augment. Our longer-term vision for this integrative project is to construct a permanent rich knowledge base that pools knowledge from many heterogeneous robot learners, as well as creating the means to add to that knowledge base. We will address understanding societal impact by informally exploring our system's impact at CMU, and on undergraduates enrolled in our new AI major.

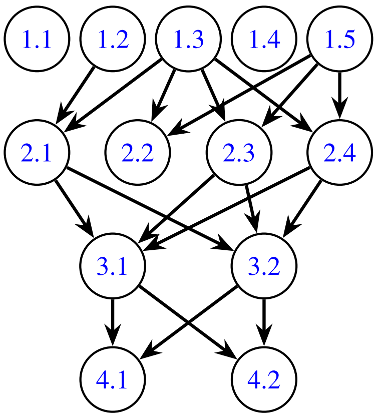
Intellectual merit: We are not the first to realize the importance of affordance and model learning for developing intelligent agents, the significance of imitation and natural language for shaping behaviors, the necessity for knowledge sharing and adaptive robotic learning. Numerous works exist in the literature on learning dynamics and/or policies [38, 88, 70, 39, 14], learning multimodal sensing [18, 50, 37, 3, 23], learning instruction to action mapping [7, 57, 58, 21, 56], learning behaviors from demonstration [44, 6, 71, 100, 64, 81, 63], and multi-robot learning [55, 41, 66, 20]. This proposal innovates in the following ways, which integrate our previous work into one system:

1) **Multimodal mapping to 3D feature maps:** The central transformative idea of the proposed research is to integrate **visual, tactile, auditory, and linguistic input** into 3D feature maps, where entities, objects and parts bind in time and have consistent object-referenced representations independent of viewpoint, as opposed to appear and disappear based on the motion of the observer or other agents or objects. The proposed representation adds a new spatial dimension to previous feature-based representations, enabling learning robots to utilize spatial reasoning such as SLAM to improve multimodal deep learning. Agents learn to see, hear, and feel, reason about temporal evolution, ground natural language, interpret, match and generalize 3D feature representations through continual collaborative learning, which guides them to attend to, focus on, and abstract important parts of the sensory streams. In this way, robots learn to imagine what is behind occlusions and results of actions and natural language goals, are able to converse and benefit from teachers' feedback through collaborative building of multimodal reasoning, and learn robot- and agent-independent representations of models, affordances, skills, and goals, supporting powerful knowledge transfer across heterogeneous robots.

2) **Closed-loop natural language teaching:** Our system will support tight realtime interaction between teachers and robots and and teachers' feedback will adjust to the competence of the learner. We expect this closed-loop feedback to accelerate policy search over open-loop instructions / action sequence pairs of previous works [4].

3) **Large-scale collaborative learning across heterogeneous robots:** Our system will support collaborative learning across heterogeneous robots and experiences, creating and using rich knowledge bases to generate and learn new behaviors. We are integrating human teaching and robot learning across a variety of robots and simulators, including a large number of low-cost robots, in contrast to the single human-single robot-single task experiments done in current research. See the Facilities section for more detail on our robots.

The team: We have assembled a team of experts from Machine Learning, Computer Vision, Robotics, and Language Understanding. PI Katerina Fragkiadaki has worked extensively on fine-grained activity understanding and visual recognition from videos by combining semantics, geometry and unsupervised learning. Co-PI Chris Atkeson has worked extensively on robot learning, manipulation, and locomotion, as well



(a) Dependency graph

Year	Research task
1	1.1 Low-cost robot development
1	1.2 Visuotactile GRNNs
1	1.3 Visual GRNNs in dynamic scenes
1	1.4 Collecting narrated demonstrations
1	1.5 Weakly supervised learning of NL semantic parser
2	2.1 Visuotactile GRNNs in dynamic scenes - model learning
2	2.2 Grounding noun phrases / referential expressions to 3D object detectors
2	2.3 Grounding NL to immediate actions
2	2.4 Grounding NL to goal imaginations
3	3.1 Open-loop NL guided visual imitation
3	3.2 Multi-robot / heterogeneous robot knowledge sharing
4	4.1 Closed-loop NL guided visual imitation
4	4.2 Continual learning / look-ahead exploration
3-4	EHT Evaluations of human-robot interaction
2-3	EHT 3D versus 2D evaluation of language grounding and model learning
2-4	EHT Task completion progress over time for the three testbeds
3-4	EHT Learning from ND versus learning from demonstrations

(b) Year - Research task

Figure 2: Timeline for the project. A dependency graph for the research tasks is shown on the left. EHT stands for evaluation-hypothesis testing. For clarity, the evaluation tasks are not shown in the dependency graph, they depend on all tasks completed thus far.

as robot design. Co-PI Wenzhen Yuan is an expert on tactile perception, both in hardware development and algorithms for understanding tactile feedback. Co-PI Tom Mitchell has extensive experience in machine learning and natural language. For example, he has recently developed methods that enable users to teach their mobile phone devices new procedures, using a combination of natural language instruction and demonstrations. Our proposal integrates Fragkiadaki's work on active vision and mobile perception [84, 27, 26], object-centric visual imitation [76], and learning from narrated demonstrations [85], Atkeson's work on robot learning [1, 10, 8, 9, 61, 62, 73, 72], learning from demonstration [12, 11, 16, 15], and tactile and auditory sensing [89, 90, 91, 92, 93], Yuan's work on tactile sensors and tactile feature learning [94, 97, 87], Mitchell's work on explanation based learning [60, 80], "never-ending" language learning [25, 59], natural language semantic parsing [49], as well as learning instructable agents [13, 52]. In what follows we describe in detail the "glue" we will create to combine our previous work into an integrated large-scale robot-human learning ecosystem.

2 Geometry-aware recurrent networks for embodied multimodal perception

This section describes our **state representations** which build upon Fragkiadaki's past work on learning to map visual features to 3D scene feature maps. The PI's lab has recently introduced geometry-aware recurrent networks (GRNNs), a family of recurrent network models **whose hidden state is a geometrically-consistent (deep) feature map of the visual scene**, and has 3 spatial dimensions paired with a feature vector, in contrast to 2 spatial dimensions of popular LSTM or convLSTM models used in the literature [42, 75, 78, 99]. We visualize GRNNs in Figure 3. The state map is updated with each new video frame in an egomotion-stabilized manner: features are transformed to cancel the (estimated) egomotion of the camera so

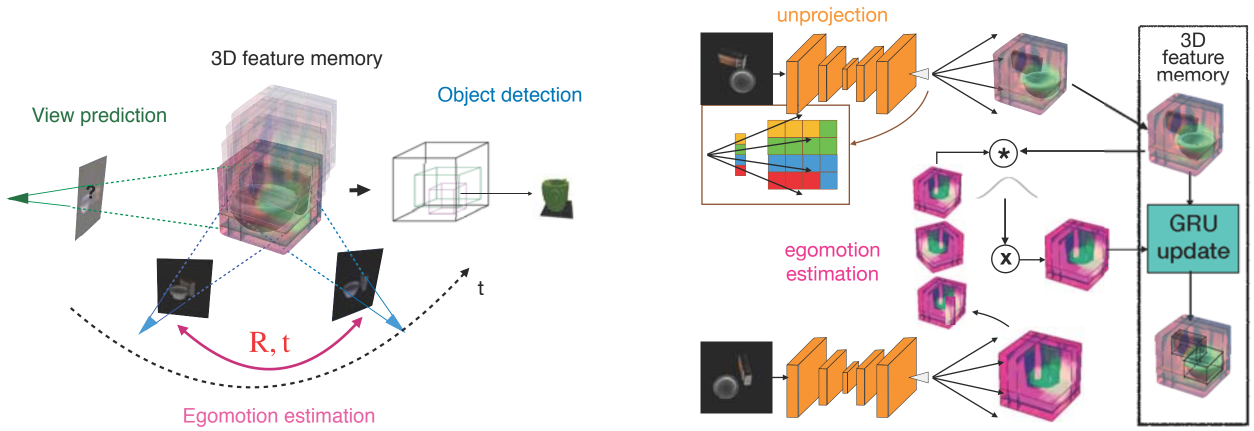


Figure 3: GRNNs 's hidden state is a multi-dimensional tensor with 3 spatial dimensions (X-Y-Z) and multiple feature dimensions, akin to a 3D map of the scene, which for every (x, y, z) grid location holds a feature vector F. The hidden state is updated with each new incoming visual frame from multiple views (left). GRNNs are equipped with differentiable geometric operations that learn to un-project 2D deep features into 3D feature tensors, estimate egomotion between a frame and the 3D feature map, stabilize against egomotion and before the state update by rotating and translating the incoming features, and project 3D feature tensors to 2D feature maps given a selected camera viewpoint (right). In this way, they learn to go back and forth between 2 dimensional (X-Y-F) sensory observations and 3 dimensional (X-Y-Z-F) spatial feature representations.

that information from 2D pixels that correspond to the same 3D physical point end up nearby in the hidden state map. Each grid feature in this 3D feature map represents information regarding a 3D physical location in the world scene. GRNNs are inspired from Simultaneous Localization and Mapping (SLAM) methods [82], but instead of point cloud maps, they build feature maps. Such features can represent a wide variety of information that is related to the downstream task, as opposed to merely 3D occupancy.

The GRNN map learns a stable model of the scene and is not affected by instantaneous object occlusions and dis-occlusions, or changes of the camera's field of view. We show in Figure 4-right 3D object detections obtained by training a 3D equivalent of MaskRCNN [40] (supervised by ground-truth 3D object boxes and 3D voxel occupancies) using the GRNN 3D feature map as input. Detected objects persist in time despite camera motion. Occluded objects that are barely or not visible in the current frame exist in the map either because they were visible in a different frame, or because our model learns to "imagine" them given a small unoccluded portion. We will use this integrated 3D deep feature map as input to the policies, dynamic models, object models, reward and affordance detectors, and any other function that uses sensory streams as input.

A long-standing debate in perceptual psychology and AI is the possible utility and role of 3D models in the form of grids, meshes, and point clouds typically found in CAD and other engineering design systems. Work in Gestalt psychology [47], 1950s artificial neural networks, and Gibsonian psychology [34] was primarily feature-based and rejected engineering-like 3D models. Much work in computer vision has focused on accurate 3D reconstruction of engineering-like models [82, 45] in terms of inferring depth maps, point clouds, 3D voxel occupancies from video data. Pointing out that **replicating the 3D world in one's head is not enough to actually make decisions**, Brooks argued for featured-based representations [22], as done by recent work in end-to-end deep learning [51] which automates learning of appropriate features. **Our proposed architectures reconcile the two sides of this debate, using feature grids of three spatial**

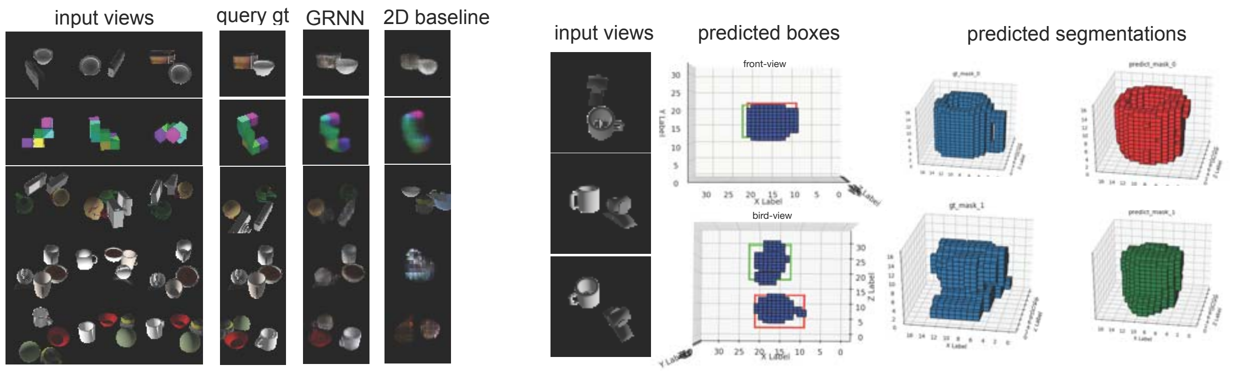


Figure 4: **View prediction (left) and object detection (right) with GRNNs.** GRNNs accurately predict views of novel scenes while 2D models fail completely to generalize, here we show the results of the Tower model of [31] (left). Furthermore, detecting and segmenting objects using the 3D feature map of GRNNs results in object detections that persist across occlusions and disocclusions (right).

dimensions as their representation.

Proposed research: visuo-tactile 3D feature mapping: We will integrate visual and tactile sensory input into our 3D feature mapping. Tactile sensors measure contact information, which usually only includes the pressure distribution, but can also include shear force, vibration, and temperature. Compared to vision, which obtains global information on a large scale, tactile sensing measures local information on the contact surface, with higher precision. Co-PI Yuan has focused on the development of a high-resolution tactile sensor called GelSight in prior work [94]. The GelSight sensor (see the Facilities section) contains a piece of soft elastomer at the surface, and uses an embedded camera to capture the change of the reflection and marker pattern of the elastomer surface. Co-PI Yuan has developed algorithms that use GelSight to estimate the hardness of arbitrary objects [98], detect slip during object grasping [95, 30], predict thickness, smoothness, fuzziness of clothing [96]. To integrate tactile measurements obtained from GelSight and other possible tactile sensors we will also place cameras in the fingers and hands to implement a proximity sense (tactile sensing at a distance). We will then integrate multi-modal observations into geometrically-consistent 3D feature maps, by taking advantage of the known finger kinematics.

Proposed research: visuo-tactile 3D feature mapping in dynamic environments: We propose to extend GRNNs to learn to model the visuo-tactile evolution of **dynamic** scenes, that is, scenes that contain moving and potentially deforming objects and textures, in addition to the motion of the observer, as a result of forces applied by the active agents. Our agent may or may not have measurements of these forces, depending on whether she is perceiving the results of her own actions or of other agents/humans. GRNNs suggest a what-where decomposition of a video scene into a content part (“what”), which is mostly constant in time, and a motion part (“where”), the egomotion of the observer and motion and deformation of the objects, which changes from frame to frame. We will allocate a separate 3D feature map to represent the appearance for each of the K moving object, in addition to the feature map of the background world scene, and a 3D motion field, a composition of a 3D rigid rotation and translation and non-rigid deformation, to represent the motion of the kth moving object, in addition to the egomotion of the observer.

We will train visuo-tactile GRNNs in static and dynamic scenes using self-supervision to predict the decomposition into object motions and additive feature changes, **simulating (imagining)** the 3D feature map of future frames, and backpropagating the error, as we detail in the next section.

3 Forward / inverse self-supervised structured model learning with GRNNs

Generalized policies and forward models of multiple temporal and spatial granularities capture *procedural* knowledge that humans and robots acquire about the world [83] through experimentation, i.e., knowledge regarding how the agent can use its cameras and end-effectors to achieve specific goals, e.g., object arrangement or sensations such as the sound of a button press. Generalized policies are functions that take as input a state and a goal representations and output (1) the action (or action distribution) that will bring the agent closer to the input goal, and (2) a termination condition that is satisfied when the goal is reached. Forward models are functions that take as input a state representation and an agent's action or action sequence and output the next state (or distribution over states) and the resulting observations. Embodied agents that move and interact with the world have access to their egomotion and actions of their end-effectors, and to the sensory (visual, auditory or tactile) outcomes of their actions. Training generalized policies and forward models to predict actions and/or action outcomes are useful forms of representation learning, often termed *self-supervised learning*, because the "labels" are provided by the embodied agent herself, as opposed to by human annotators. **The main research challenge is to build generalized policies and forward models whose state representations strongly generalize across environment variations**. Despite progress, such generalization has not been seen in the literature [88, 70, 39, 14]. Some approaches use toy 2D worlds where objects cannot occlude one another [88, 14], some do not model object appearance assuming the same object will be encountered at test and training time [2, 32], and some model full frame feature encodings, without trying to compute scene structure [79], and cannot easily generalize to novel situations. Our conjecture is that the proposed multimodal 3D feature representation holds promise for such generalization.

Both the state and goal representations for our forward models and generalized policies will be 3D multimodal feature tensors produced by GRNNs, of different spatial resolutions. PI Fragkiadaki's lab has trained forward models for visual GRNNs that predict the results of egomotion of the agent using a short video frame sequence as input. Results are shown in Figure 4-left. Even when GRNNs are trained in scenes that contain two objects and are tested on scenes with four objects, they effectively can "imagine" how scenes with four objects look like from different viewpoints, that is, **the visual forward model generalizes effectively to truly novel scenes.** In contrast, geometry-unaware 2D models fail to generalize (Figure 4-left). Our conjecture is that exploiting 3D and object-centric priors, in place of plain 2D convolutional state encoding of the majority used in previous research [79], we will be able to generalize better, as supported by our preliminary results on predicting results of egomotion.

Proposed research: Learning structured visuo-tactile 3D forward models and generalized policies: We will train visuo-tactile GRNNs that given a current visuo-tactile streams and the action stream of the agent they predict both visual and tactile feedback, extending our work on view prediction. Our approach will optimize over moving object and object part detections, object and camera motions, additive feature changes and background scene appearance via a combination of gradient-based learning and reinforcement learning, using pretraining of object detectors in simulation, as well as curriculum learning from rigid to non-rigid scenes, to assist with bad local minima. Since predicting raw tactile input or images may be noisy, we will use instead a discriminative alternative: we will predict intermediate low-level feature embeddings, and train them to better match (have smaller Euclidean distance to) the embeddings extracted from the corresponding (future) sensory inputs than other non-corresponding sensory inputs (i.e., metric learning). We will further use losses that exploit **cross-modal temporal coincidence of visual, tactile and auditory** sensations: embeddings of temporally coincident visual, tactile and auditory inputs should be placed close in the embedding space, while embeddings of non-temporally coincidental visual, tactile and auditory sensations should be placed far from each other. Such cross-modal metric learning has been explored in previous work by co-PI Yuan [87, 24]. Here, we propose utilizing it for training 3D feature maps, as opposed to 1D embeddings. In both forward models and generalized policies we will use latent variable to handle stochasticity of actions and states, as we have done in our previous work with different network architectures [33].

Mental practice: The proposed research will produce multimodal forward models and generalized policies that, given visuo-tactile and auditory streams, will be able to infer the underlying 3D physical reality, which objects move and how, robust to occlusions from the agent's hands or other objects, as well as simulate (imagine) such reality forward in time under different behaviors. By comparing the results of our forward models against the goal representations, we will search over the right actions to choose, either using gradient descent or evolutionary methods [32, 70].

4 Collaborative model learning via imitation and instruction

We described training of state representations by predicting sensory outcomes (forward models) or actions that cause sensory transitions (generalized policies). What policies should our agent use to collect data to fit such models? Curiosity-driven exploring agents [65, 54, 74] seek to increase their surprise and learning progress by focusing their interactions on the yet unknown parts of the state space. Curiosity-driven exploration in the real world has the following severe limitations: a) It may take the curious agent **a very long time** to figure out how to reach interesting states of the environment, e.g., to successfully operate an elevator button. Sample efficiency is crucial for artificial agents because their end-effectors are much more fragile and much less agile than that of humans: they wear and break, and they are slow. b) It is **unsafe** to explore and acquire knowledge in a human environment with partially trained or untrained policies. c) The knowledge acquired by our agent **may not be easily controllable** by humans since the agent does not know the mapping of his/her state representations to natural language. d) **The agents will not be able to learn about human reward functions**, e.g., that when gripping a glass of water, it is preferable to not touch the inside of the glass.

Humans do not learn from scratch driven solely by their curiosity. They use imitation and natural language (NL) to share each other's procedural knowledge about the world. We thus propose collaborative human-robot interactions for learning models, policies, and affordances, through open-loop and closed-loop human teaching. Human teachers guide the robots through visual or kinesthetic demonstrations, and they concurrently verbally describe objects, actions, goals, or mistakes. During closed-loop teaching, the human can provide visual, kinesthetic, and verbal feedback while watching the robot perform a task. We use the term narrated demonstrations (ND) to refer to visual or kinesthetic demonstrations paired with natural language. Input from human teachers can assist model learning in the following three ways:

1) Causality and attention: NL and gesticulation of the teachers suggest the important objects or parts of the scene to attend to, and clarify the intention of a particular demonstration. The attended objects and parts become the nodes in our proposed graph neural network (GNN) reward detectors, which abstract away from (forget) the rest of the scene. The less unnecessary information supplied to a classifier or regressor, the better it generalizes in the future. Instead of relying on a large number of demonstrations to learn to attend to the right features, teachers directly supervise such attention through gesticulation and NL.

2) NL grounding injects critical semantic information for state representation learning: The robot learns to detect objects, attributes, actions, goal completion, and referential expressions in their multimodal 3D feature representations. NL supervision is critical when multimodal self-supervised learning does not suffice to capture important states or state changes. For example, while a switch occupies a very small number of pixels and its state (ON or OFF) is not easily detectable through self-supervised metric-learning (as described in sections 2,3), natural language grounding of the description *"now the switch is off"* trains an ON/OFF attribute classifier over 3D multimodal features extracted from the switch and makes them sensitive to this state change.

3) Accelerating safe exploration: Closed-loop teaching helps the agent discover the right actions to carry our the task, while preventing her from unsafe operations, using basic NL grounding of actions, e.g.,"Stop!".

Proposed research: Collecting narrated kinesthetic and visual demonstrations: We will collect a dataset of *narrated visual and kinesthetic demonstrations* in virtual and real environments. Human teachers equipped with microphones will name objects in the scene, describe their relationships, indicate the activities being performed, explain the outcomes, and, gesticulate deliberately so as to guide the learner towards the correct interpretation of the natural language description. Verbal narrations will be automatically transcribed into textual descriptions using the Google speech recognition API [43]. Errors made during speech recognition, which are rare, will be corrected by hand. The synchronization of the narration to the video, along with present-tense descriptions, provide a natural alignment of the semantic content to the visual stream. Consecutive demonstrations are easily temporally segmented by considering their alignment to natural language utterances. The scalability in terms of human effort of verbal narrations far surpasses that of video post-transcription [68] or detailed scene graph annotations [48], considered in previous works.

The more instrumented the environment, the better the world state is observed and the easier the visual recognition, natural language interpretation, and action inference for a particular demonstration. We can use the extra information from an instrumented environment to train policies that only use a limited set of sensors available in their operating environment, as well as capture richer narrations and gestures. For example, in Virtual Reality the state of the world (objects, their pose, their attributes, etc.) is fully observed. This allows **easier grounding of natural language utterances to disentangled feature representations**, e.g., speed, pose, spatial location etc. of the objects in the environment. Our plan is to train natural language grounding using narrated demonstrations in heavily instrumented environments first, and then proceed to less instrumented ones. PI Fragkiadaki's lab has already started such data collection, as we describe in the Facilities section. We will use paraphrasing to augment the instructions using Amazon Mechanical Turk, inspired by [17], in order to handle natural language variability.

Proposed research: From narrated demonstrations to neural-symbolic forward models and policies: *Natural language semantic parsing and grounding:* During open and closed loop teaching, teachers utter natural language instructions and descriptions that have a rich set of functionalities, far beyond the object category labelling of Imagenet and COCO annotations [29, 53]: Utterances may refer to entities in the environment, e.g., *"the black puppy"*, *"the red mug next to the bowl behind the orange"*, describe the state of the world or the results of the actions, a.k.a. post-conditions, e.g., *"the window is open"*, *"the brown mug is larger than the green one"*, *"the screen is brighter than before"*, describe the actions that are taking place (descriptions), or ought to take place (instructions), e.g., *"I am taking the block out of the bucket"*, *"press the button"*, *"lift it higher"*, *"move slower"*, *"look more to the left"* or provide information directly in terms of pre-conditions/actions/post-condition, e.g., *"if there is smoke coming out of the oven, I switch it off"*.

How should we ground this rich natural language to facilitate policy and model learning? In our recent work [85], depicted in Figure 5, we made the first attempt to use narrated video demonstrations of pick-and-place activities to learn pick-and-place policies instructable by natural language. We used the NDs to learn **perceptual reward detectors** to detect in images desired post-conditions (e.g., "the coca cola should be inside the wooden box") and used those to guide policy learning, replacing manually coded rewards. Our main realization was the necessity for the reward detectors to operate in a 3D attributed space, in order to be accurate but also to generalize across viewpoints, and to provide shaped—instead of binary—rewards to the policy search method. Indeed, state of the art policy learning methods [5, 67] in simulation or in the real world assume 3D object centroids of goal configurations. We thus propose right below **grounding post-conditions / goals / rewards to goal detectors in the space of our 3D multimodal feature tensors**.

We will use Multiple Instance Learning and the temporal synchrony of narration and video to jointly learn (1) neural sequence models that parse NL utterances into structured attributed intention forms, as we did with referential expressions in [85], and (2) to ground NL structured attributed intention forms to the corresponding object, attribute, scene and action detectors, as follows:

1) We will ground **noun phrases to modular detector programs that detect the corresponding referents** in our 3D multimodal feature scene representations. We will update accordingly our modular detectors with those inferred associations, or **instantiate novel detectors** when encountering novel categories. In this way, our detector vocabulary will grow with experience. We will avoid catastrophic forgetting [46] by progressively growing the capacity of our detectors, similar to progressive nets [69].

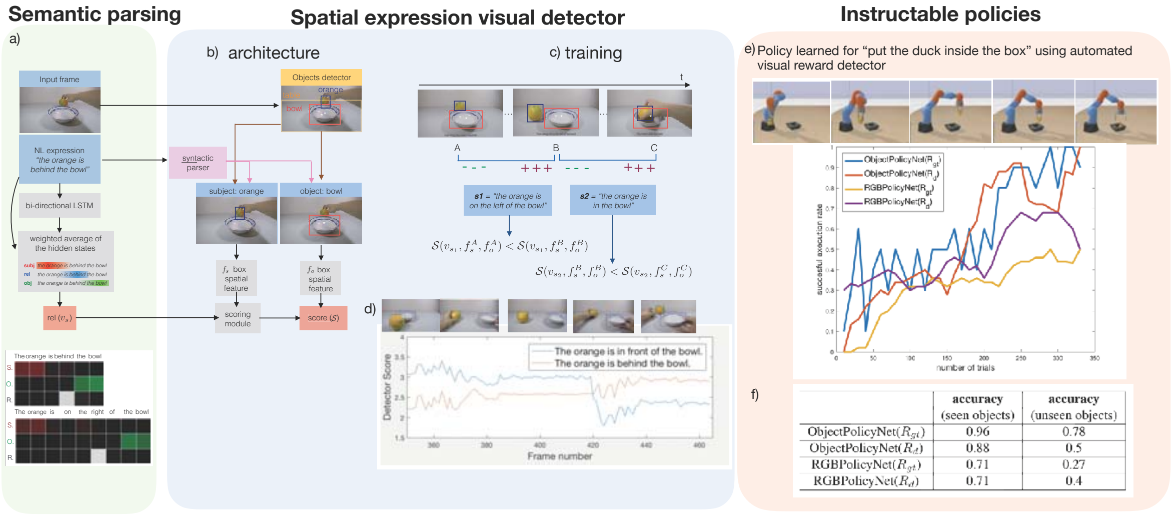
2) We will ground **goal NL descriptions to goal sensory feature imaginations**, namely, 3D feature representations of the desired resulting state, as well as the corresponding objects and attributes associated with it. This extends our NL reward learning work [85] to operate on a 3D multimodal feature space. We will further train reward detectors, that given the current and goal state, will provide a measure of distance to guide policy learning. Such reward detectors are graph neural networks on top of 3D object/part/point detections on our 3D feature tensors.

3) We will ground **actions to robot's modular visuomotor programs** assembled using perceptual modules, such as, detectors, and motor modules, i.e., policies, e.g., "look left" or "focus on the red object". We pretrain such language to motor program mapping using narrated demonstrations in virtual reality. Following our earlier work on instructable mobile agents [13], our semantic parser will begin with a primitive lexicon and language capability that enables users to refer to each of the primitive robot sensors and effectors, so that users can teach more complex visuo-motor procedures, sensing procedures and subgoal states grounded in terms of these primitives, making them directly executable by the robot.

4) We will ground facts about the world state to updates of high level symbolic models of the environment regarding object similarities, ownership relationships, person-place and object-place relationships, etc.

Prior Work - Semantic parsing of mobile phone instructions: The research proposed here will build our recent research into interactive instruction of mobile phone devices using natural language and demonstrations. In that work, we view the mobile phone as a robot which contains both physical sensors and effectors (e.g., microphones and sound alarms) as well as cyber sensors and effectors (e.g., calendar and email readers and writers). In that work we have developed a prototype agent, LIA, that enables users to teach their phones new procedures through natural language instruction [13]. For example, a user may say to their phone "tell Katerina that I'm arriving soon." The phone (LIA) then responds "I don't understand, do you want to teach me?" and the user may then say "Create a new email, put Katerina's email address in the recipients list. Then put "I'm arriving soon" in the subject field, and send it." As a result of this training episode, LIA learns two types of knowledge: first, it learns the detailed steps of the procedure which the user had in mind, i.e., the motor program; second, it learns an improved natural language competence enabling it to now parse commands such as "Tell X that Y" into semantic expressions grounded in the primitive sensor-effect capabilities of the agent (the phone in this case).

Proposed work: neural-symbolic NL guided visual imitation: During real-time NDs, the student and teacher share the same virtual or physical workspace and the student actively observes the teacher's actions, akin to a child - parent interaction. This setup (1) facilitates inference of the teacher's actions because the student-robot will choose its preferred camera viewpoint to observe the teacher's actions over time, (2) can take place during deployment in a human environment, outside the lab. We will build upon PI Fragkiadaki's recent work on object-centric visual imitation [76], as well as her work on active vision for recognition and manipulation [26]. In [76], two graphs are instantiated, one for the student and one for the teacher, where nodes represent objects or object parts and their attributes, and edges represent cross-node 3D spatial relationships. The nodes between imitator and teacher graphs are in one-to-one correspondence. **The objective of imitation is to match the edge and object attributes across the two graphs**, and it is optimized with


Semantic parsing
Spatial expression visual detector
Instructable policies
a)
b) architecture
c) training
e) Policy learned for "put the duck inside the box" using automated visual reward detector
Input frame
NL expression "the orange is behind the bowl"
bi-directional LSTM
weighted average of the hidden states
subj
rel
obj
rel (v_s)
syntactic parser
Objects detector
orange
bowl
subject: orange
object: bowl
f_s box spatial feature
f_o box spatial feature
scoring module
score (S)
The orange is behind the bowl
The orange is on the right of the bowl
S
O
R
t
A
B
C
- - -
+ + +
s1 = "the orange is on the left of the bowl"
s2 = "the orange is in the bowl"
$\mathcal{S}(v_{s_1}, f_s^A, f_o^A) < \mathcal{S}(v_{s_1}, f_s^B, f_o^B)$
$\mathcal{S}(v_{s_2}, f_s^B, f_o^B) < \mathcal{S}(v_{s_2}, f_s^C, f_o^C)$
d)
The orange is in front of the bowl.
The orange is behind the bowl.
Detector Score
Frame number
ObjectPolicyNet(R_{gt})
ObjectPolicyNet(R_d)
RGBPolicyNet(R_{gt})
RGBPolicyNet(R_d)
successful execution rate
number of trials
f)

	accuracy (seen objects)	**accuracy** (unseen objects)
ObjectPolicyNet(R_{gt})	0.96	0.78
ObjectPolicyNet(R_d)	0.88	0.5
RGBPolicyNet(R_{gt})	0.71	0.27
RGBPolicyNet(R_d)	0.71	0.4

Figure 5: **Learning instructable pick-and-place policies from narrated video demonstrations [85]:** The NL expression is parsed by a bidirectional LSTM to localize subject, objects and relationship subphrases, and produce corresponding word vectors. Given the semantic parse and an RGB image, a reward detector returns a detection score S based on how well the image matches the natural language utterance. The reward detector is trained with weakly supervised metric learning (c). Given a video sequence, our reward detector effectively detects changes of spatial configurations for the depicted objects (d). We use the learned reward to train a manipulation policy to achieve the corresponding spatial configuration in simulation (e). Modular, object-factorized neural architectures both for the reward detector and the policy network (*ObjectPolicyNet*), generalize better than policy networks that use the whole RGB image as input (*RGBPolicyNet*) (f).

dynamic programming using a linear quadratic regulator (LQR), by learning local models online with random exploration. All imitation methods that reply on 2D object detections are sensitive to occlusions. Yet, occlusions due to moving hands are parsimonious in manipulation. We propose, in place of 2D detectors, to instead detect and track object and object parts in our 3D feature maps, extracted from executor's and imitator's environment using GRNNs and NL noun phrases. The dynamic visuotactile GRNNs of Section 2 represent in their hidden state the full state of the environment, unaffected from occlusions. We will train active vision policies that learn to move the robot's camera to further facilitate inference of the teacher's activity and thus its imitation. Once pretrained with real-time active visual imitation, we will train our robots from internet videos, building upon their visual competence for human activity understanding.

5 Evaluation

We will implement and evaluate three instantiations of our learning ecosystem: **Toys/Kits Testbed:** Multiple small low-cost robots working with educational toys and kits. **Deformable/Liquid/Granular Testbed:** Human-scale robots working with deformable, liquid, and granular materials as is found in food preparation and science experiment kits for children. **Social/Affordance Testbed:** Ubiquitous human-scale co-robots (cobots) learning about social interaction and object affordances by interacting with people in our work environment.

Toys/kits Testbed: Educational toys and kits are designed to stimulate and facilitate perceptual, affordance, model, skill, and cognitive learning. Toys for infants typically demonstrate numerous manipulation skills and affordances (e.g., knobs to turn, buttons to press), are safe to operate, can break and be replaced easily and inexpensively, and use exaggerated visual and audio cues to guide the learner. Educational K-12 toys and kits we will focus on include construction toys and kits such as Lego “Chain Reactions”, Keva “Contraptions”, Marble runs, Rube Goldberg kits, and Jenga. We will also focus on learning to play simple musical instruments such as keyboards, drums, rattles, and xylophones to emphasize multimodal visual, aural, and tactile learning. More details of our prior work in this domain are provided in the Facilities section. In this testbed, we will use human-scale robots with complex multi-fingered hands, as well as low cost robots with simple grippers which we will equip with low cost handheld tools such as solenoids for snapping together Lego parts and for “shooting” balls in small-scale versions of billiards, croquet, and mini (put-put) golf. We will construct virtual reality versions of “rigid body” toys as well as our robots. Fun videogame-like simulations will be used to engage a large number of human teachers on the web for virtual teaching, and this domain is an excellent vehicle for outreach to schools and museums. Evaluation in this testbed will focus on performance: how well can robots perform activities suggested by the instructional materials at varying levels of detail, repair broken constructed setups, and construct setups that achieve new task specifications. For specified construction tasks or in learning from demonstration, objective measures based on motion capture and other forms of object (typically a ball) tracking will provide objective measures of performance. Experimenters and independent judges (such as Mechanical Turk workers) will provide subjective performance measures by grading how well constructed setups match task specifications or demonstrations. In the case of musical instruments, the sounds generated will be objectively scored against ideal versions created by humans, as well as being subjectively rated by experimenters and independent listeners. Experimenters and also independent humans will introduce flaws in constructed setups to evaluate robot performance on repair tasks. We will also use suggested projects from the kits and instructional materials as tasks to perform without detailed instructions.

Deformable/Liquid/Granular Testbed: We choose deformable, liquid, and granular materials manipulation to explore domains where modeling the task is more difficult and cultural knowledge of how to do tasks is more important than model-based planning or reasoning from scratch using idealized models of the underlying physics. We have chosen food preparation as a good domain, partly because we are already working on food preparation in collaboration with Sony. We have chosen making salads as an initial focus task because there is an existing infrastructure and progression of difficulty: one can start with “kit/bag/box salads” where all items necessary can be purchased pre-washed and pre-cut in a bag or box, progress to working with bulk pre-cut materials, and graduate to preparing a salad from whole vegetables. Similarly, with baking we can progress from baking kits for children to cookie, cake, and muffin mixes commonly found in supermarkets, to preparing materials “from scratch”. Science experiment kits for children allow robots to explore mechanics, chemistry, and electricity while performing manipulations expected of young children. Food preparation and science experiments will also be learned from online instructions, recipes, and videos, which opens up a huge range of instructional material. More details of our work in these domains are provided in the Facilities section. In addition to the evaluation methods and statistical analyses of results previously discussed, food preparation needs to be evaluated subjectively by human tasters as well as being judged more objectively for appearance and mechanical properties. Science experiments for children can typically be objectively evaluated by experimenters and independent judges in terms of obtaining the expected result, as well as subjectively evaluated for how the various sub-tasks were performed.

Social/Affordance Testbed: Our third domain, human-scale co-robots (cobots), builds on existing co-robot infrastructure created by Professor Manuela Veloso and her PhD students [19, 86], who supports the use of the cobots for our third testbed. For details on the cobot infrastructure, please see the Facilities section. We will add simple arms and hands to the existing cobots. Our goal will be to enhance existing human-cobot interactions to teach the cobots social rules and etiquette while acting as tour guides, meeting and seminar hosts as well as general hosts, companions and other forms of social facilitators, mail and food deliverers, vendors, patrollers, monitors, cleaners, entertainers, and trash and recycling system managers. A concrete goal is implementing a functioning trash and recycling management system with sorting and correcting extensive human error and non-compliance. Robots will also learn about affordances of everyday objects such as doors, elevators, furniture (especially mobile tables and chairs with adjustment controls), books, backpacks, desktop, laptop, and tablet computers, phones, sinks, water fountains, refrigerators and other kitchen objects, ice machines, lights, thermostats, whiteboards, black boards, markers, chalk, erasers, food service objects (e.g., containers, napkins, utensils, cups) mops, brooms, and other cleaning equipment, toilets, towel dispensors, hand dryers, packages, package material waste, various forms of recycling and trash infrastructure, and A/V equipment (it is our hope that our meetings and seminars will run more smoothly if a robot is in charge). We will emphasize a variety of manipulation skills, such as object pick-and-place, pushing, pouring, stacking, cleaning, etc. Cobots are already successfully navigating our building while executing a variety of prespecified tasks, including parcel delivery and visitor escort. More details of our work in this domain are provided in the Facilities section. The Facilities section also describes our instrumented environments available for fine scale human-robot behavior capture, which augments learning from ubiquitous human-robot interactions. Robot behavior in the Social/Affordance Testbed can be evaluated for performance. Experimenters and independent judges can rate whether the correct behavior was selected, and how well that behavior was performed. In addition, we can test what was learned by using it to narrate or explain human or robot behavior in similar situations.

Evaluation, comparison, and hypothesis testing: Our work will include both testing of scientific hypotheses and engineering development of design methods, architectures, and algorithms that facilitate building a rich knowledge base and enable multiple heterogeneous robots and teams of robots and humans to learn in a unified system. We will evaluate our never-ending affordance learners on their ability to improve over time, as well as their rate of improvement, i.e., their ability to learn how to learn better in the future, as stated in our requirements for successful never-ending learning architectures [59]. In all three testbeds, we will evaluate our agent's abilities in the following tasks:

Affordance and skill learning: For each testbed, we will create a task benchmark data set containing 100 tasks per testbed. These are tasks whose variations have been demonstated to our agents during training. Periodically, during training, we will evaluate our agents on a subset of 10 tasks, where performance is judged by human scorers. Note that our agents would have learned their own reward detectors for the tasks, but we cannot use those as ground truth since they may be erroneous. In general, statistical analyses will be straightforward (N out of M test tasks were successfully completed giving a success rate of Z%). Part of the proposed research will be to develop metrics that measure partial performance and learning progress.

Natural language understanding: We will benchmark the ability of our agents to understand and carry out human instructions in particular visual and execution contexts in two ways: 1) we will use a benchmark set of instruction-to-task execution mappings collected using narrated demonstrations in virtual reality and record task completion rates throughout the training period of our agents. 2) we will conduct studies at regular time intervals to quantify the satisfaction/frustration of our teachers with the responsiveness of the student-agents, both those accessible through the web as well as those sharing the same workspace, using appropriate questionnaires. We will use their feedback to improve our pretraining open-loop teaching stage as well as closed-loop learning algorithms.

Visual imitation: Given a novel (potentially silent) demonstration of a novel activity, we will evaluate how fast our agents learn to perform the depicted task. We will create a benchmark of demonstrations by varying their novelty against the training set encountered by our agents.

This proposal puts forward a set of hypotheses regarding architectural and supervision design choices claimed to be important for behaviour and affordance learning. We state them below for clarity and describe tests for their verification or disproof.

Does temporal integration of visual and tactile sensations into 3D feature maps lead to improved

affordance recognition / policy learning / language grounding? We will compare policies learned by varying the perceptual front-end between using conventional frame-stacked visuotactile features, traditional engineering 3D representations such as pointclouds and the proposed 3D feature maps, and evaluate generalization of the skill policies learned. We expect that learning based on timing of event occurrences (associational learning), and language about temporal relationships will not be improved by mapping to three dimensions, but spatial learning and learning from language about spatial relationships will be greatly enhanced by the use of 3D features. We expect learning using the proposed 3D features will be faster and perform better than learning based on traditional 3D representations or 2D feature maps. Engineering challenges include managing GPU memory for such memory-intensive representations.

Whether and how much closed-loop teaching helps over open-loop narrated demonstrations, and whether and how much learning from NDs helps over learning from demonstrations? Here we will evaluate robot learning (task completion) from the same training data with and without closed-loop teachers' feedback, and with and without human narration.

Whether and how much predictive models accelerate acquisition of manipulation skills in imitation and reinforcement learning? We will evaluate the learned dynamics models in model predictive control, by unrolling them forward in time [70] for action selection, or exploration in the training control loop [2].

Other issues and hypotheses we would like to explore include: a) A special simplified language (Robot Esperanto) will improve human-robot interaction. accelerates learning and improves performance. c) Multi-level feature representations including symbolic information can improve learning. d) Our approach actually reduces the cost of robot programming, rather than just changing the form of robot programming.

6 Broader Impacts:

Impact on society: We expect to make programming robots cheaper, and make robots more useful, particularly for domestic and care robots supporting everyday life activities and unstructured activities such as cleaning and repair. Programming a robot to do a desired task is difficult and expensive (typically requiring one graduate-student-year for state of the art dynamic tasks). Adding the necessary error handling is many times more expensive, as it is very difficult to anticipate all the things that will go wrong. We learned from our participation in the DARPA Robotics Challenge (DRC) that **designing robust behaviors for robots is very difficult, even for professionals.** Small changes in the task caused robots to fail, and even to fall or crash. We propose methods that will help automate robot programming and error prevention and handling. We want to **enable robot workers and explorers to make simple plans and solve minor problems autonomously,** and be able to **attain a safe state and ask for help when major errors or problems occur.** Endowing robots with the ability to solve simple problems, learn, and ask for help when needed is more cost-effective than trying to make robot programs free of bugs and conceptual errors.

Impact on the research community: We expect to continue to make demonstration and robot data available on the web. We have had great success making public most data collected in our Motion Capture Lab (mocap.cs.cmu.edu and kitchen.cs.cmu.edu) and Panoptic Studio (domedb.perception.cs.cmu.edu). Data made available so far has been acknowledged in several hundred papers, mostly from the computer graphics, animation, and vision communities worldwide. This form of usage is freely available to all, including those from non-Ph.D. and/or minority-serving institutions. We will host visitors who wish to use our facilities, as we do now. Our technologies are being shared by being published, and papers and software will be available electronically. We will maintain a public website to freely share our demonstrations and robot data with additional video material. We will present our work at conferences and publish it in journals, and will use these vehicles to advertise our work to potential collaborators in science and industry.

For a more complete description of our *Dissemination Plan,* please see our Data Management Plan.

Outreach: We have two outreach efforts aimed at reaching a wide audience. The first builds on our experience with Disney, in which our soft robotics work inspired the soft care robot Baymax in the Oscar-winning Disney movie *Big Hero 6*. We will coordinate with the ongoing "Big Hero 6" Disney TV show and Disney park activities, and publicize our effort as "Building Baymax". A second wide audience effort is our effort to create a robot museum. Our work on low cost robots will be used as the basis of example museum exhibits and school activities.

Broadening Participation: In terms of more general outreach to under-served populations, we will make use of ongoing efforts in the Robotics Institute and CMU-wide. These efforts include supporting minority visits to CMU, recruiting at various conferences and educational institutions, and providing minority fellowships. As the Robotics Institute PhD admissions chair in 2016, Atkeson led a process which resulted in 31% of acceptances going to female applicants. As a member of the Robotics Institute faculty hiring committee in 2017, Atkeson participated in a process that led to approximately half the interviewees being female. Half of the faculty hired were women. As the head of Robotics Institute hiring in 2018, Atkeson led a process in which again approximately half the interviewees were female, and 3 out of the 4 hires were female. In the last few years we have tripled the number of women faculty in the Robotics Institute (3 to 9). Atkeson is assisting efforts at CMU to raise money for fellowships for students who can help us in our efforts to serve diverse populations and communities, including our own. Fragkiadaki organizes the CMU chapter of the AI4ALL national program, with the first version presented in July 2018: a three week program for 20 high school students from disadvantaged local schools, to expose them to the excitement of AI, its potential societal impact, and what people do in college, graduate school, and beyond. The first instantiation went very well, and both the participants and the organizers learned a lot. Preliminary results of the proposed research were presented to the students, who were very enthusiastic. At the end of the program, they all indicated in surveys that they wanted to pursue AI as their field of college studies. We believe the research proposed in this proposal is intuitive enough to excite young students, and we plan to organize one project in the AI4ALL school of summer 2019 on visual and tactile recognition by robotic agents. Currently PI Fragkiadaki is mentoring three Ph.D. students, one of which is female and leads the research on GRNNs. A recently graduated undergraduate student from PI Fragkiadaki's lab, Ricson Cheng, was selected as **a runner-up for the CRA outstanding undergraduate awards in 2018** for his work on active vision and geometry-aware RNNs.

Technology Transfer: The best way to transfer technology is by having students go to industry. Three recent students work at Boston Dynamics transferring our work in robotics to commercial applications, one recent student and recent postdoc work on self-driving cars at Uber, one recent student works on self-driving cars at Apple, and one recent student works on humanoid robotics at the Toyota Research Institute. An older former student is the CTO of the Amazon drone effort. Several older former students work at Google. We are thrilled that we and our students are part of the robotics revolution.

Education and Curriculum Development Activities: We intend to infuse this research and the robot experimental program into our new AI major at CMU, especially through project courses and student theses. We will develop course material on robot learning and reasoning, which will be influenced by our research and freely available on the web. The PIs currently teach several courses that will benefit from this material. For example, *10-403 (undergraduate) and 10-703 (graduate): Deep Reinforcement Learning and Control*, *10-898: Language Grounding to Vision and Control*, and *16-745: Optimal Control and Reinforcement Learning* directly address the research areas in which this proposal is embedded. We also teach a course designed to attract undergraduates into the field, *16-264: Humanoids*, which is currently using the first version of the **Toys/Kits Testbed.**

7 Results from Prior NSF Support

The most relevant recent award for Atkeson is: *(a) NSF award:* IIS-1717066 (PI: Atkeson); *amount:* $440,000; *period:* 8/1/17 - 7/31/20. *(b) Title:* RI: Small: Optical Skin For Robots: Tactile Sensing and Whole Body Vision *(c) Summary of Results:* This recent grant is supporting work on developing optical approaches for tactile sensing as well as whole body vision (eyeballs all over the body). We will develop robot hands that complement the robot skin.

Intellectual Merit: This project will enable robots to feel what they touch. The key idea is to put cameras inside the body of the robot, looking outward at the robot skin as it deforms, and also through the robot skin to see nearby objects as they are contacted or avoided. We turn tactile sensing into a computer vision problem, taking advantage of recent progress in computer vision. This approach addresses several challenges: 1) achieving close to human resolution (a million biological sensors) using millions of pixels, 2) reducing occlusion during grasping and manipulation, and detecting obstacles before impact, and 3) protecting expensive electronics and wiring while allowing replacement of worn out or damaged inexpensive skin. Technical goals for the project include first building and then installing on a robot a network of about 100 off-the-shelf small cameras (less than 1 cubic centimeter) that is capable of collecting information, deciding what video streams to pay attention to, and processing the video streams in real time to estimate forces, slip, and object shape. A transformative idea is to aggressively distribute high resolution imaging over the entire robot body. This reduces occlusion, a major issue in perception for manipulation. Building a camera network of hundreds of cameras on a mobile skin, and building a multi-modal sensing skin, is synergistic with developing the proposed system.

Broader Impacts: Robots with better sensing can more safely help people. **Development of Human Resources:** The project involves one graduate student. We have weekly individual meetings and weekly lab meetings. The graduate student is performing research, making presentations to our group, and will give conference presentations and lectures in courses. We will put the graduate student in a position to be a success in academia and industry.

(d) Publications resulting from this NSF award: [93, 28]. *(e) Other research products:* We have made instructions on how to build our tactile sensors available on the web. *(f) Renewed support.* This proposal is not for renewed support.

Tom Mitchell has been a PI or co-PI on 15 NSF grants over the years, and on three recent NSF grants that have focused on developing and analyzing machine learning algorithms for large scale continuous learning. He was co-PI on NSF IIS1250956 “BIGDATA: Small: Big data for everyone,” ($548,417, Aug 2013 to May 2017). **Intellectual merit:** Developed a continuously learning system that can be retargeted to extract knowledge from the internet in different domains, leading to advances in self-reflection [Platanios 2014; 2017] and information extraction [Saparov 2017]. **Broader impact:** This research seeks to bring the benefits of learning from internet data to new fields, and resulted in support for two Ph.D. advisees. Mitchell was co-PI on NSF IIS1247489 “BIGDATA: Mid-Scale: DA: Collaborative Research: Big Tensor Mining: Theory, Scalable Algorithms and Applications, ($894,892, Dec 2012 to Nov 2017). **Intellectual merit:** Improved scalability for tensor and coupled tensor/matrix factorization, driven by real-world applications [Papalexakis 2014; 2016], [Xiao 2017]. **Broader impact:** More scalable tensor-based algorithms have impacted big data analyses, including in cognitive neuroscience. This has supported one of Mitchell’s graduated Ph.D. student and one female post-doctoral researcher. Mitchell is currently co-PI on NSF 1535967 ”AitF: FULL: From Worst-Case to Realistic-Case Analysis for Large Scale Machine Learning Algorithms.” **Intellectual merit:** This research seeks new theory to characterize real-world machine learning, and has resulted in understanding new ways to use unlabeled data to estimate learning accuracy [Platanios 2016]. **Broader impact:** This has supported one of Mitchell’s PhD advisees.

There is no prior NSF support for Katerina Fragkiadaki or Wenzhen Yuan.

References Cited

[1] *E. Aboaf, S. Drucker, and C. Atkeson. Task-level robot learning: Juggling a tennis ball more accurately. In *IEEE International Conference on Robotics and Automation*, pages 1290–1295, Scottsdale, AZ, 1989.

[2] Arpit Agarwal, Katharina Muelling, and Katerina Fragkiadaki. Skill-guided look-ahead exploration for reinforcement learning of manipulation policies. In *Association for the Advances of Artificial Intelligence (AAAI)*, 2018.

[3] Peter K. Allen, Andrew T. Miller, Paul Y. Oh, and Brian S. Leibowitz. Integration of vision, force and tactile sensing for grasping. *Int. J. Intelligent Machines*, 4:129–149, 1999.

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *CoRR*, abs/1711.07280, 2017.

[5] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *CoRR*, abs/1707.01495, 2017.

[6] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robot. Auton. Syst.*, 57(5):469–483, May 2009.

[7] Yoav Artzi and Luke Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*, 1:49–62, 2013.

[8] C. Atkeson and J. Santamaria. A comparison of direct and model-based reinforcement learning. In *International Conference on Robotics and Automation*, 1997.

[9] *C. G. Atkeson. Nonparametric model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 10, pages 1008–1014. MIT Press, Cambridge, MA, 1998.

[10] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

[11] C. G. Atkeson and S. Schaal. Learning tasks from a single demonstration. In *Proceedings of the 1997 IEEE International Conference on Robotics and Automation (ICRA97)*, pages 1706–1712, 1997.

[12] C. G. Atkeson and Stefan Schaal. Robot learning from demonstration. In *Proc. 14th International Conference on Machine Learning*, pages 12–20. Morgan Kaufmann, 1997.

[13] Amos Azaria, Jayant Krishnamurthy, and Tom M. Mitchell. Instructable intelligent personal agent. *Proceedings of the AAAI Conference*, 2016.

[14] Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *CoRR*, abs/1612.00222, 2016.

[15] *D. C. Bentivegna, C. G. Atkeson, A. Ude, and G. Cheng. Learning tasks from observation and practice. *Robotics and Autonomous Systems*, 47:163–169, 2004.

[16] *D. C. Bentivegna, C. G. Atkeson, A. Ude, and G. Cheng. Learning to act from observation and practice. *International Journal of Humanoid Robotics*, 1(4):585–611, 2004.

[17] Jonathan Berant and Percy Liang. Semantic parsing via paraphrasing. In *ACL (1)*, pages 1415–1425. The Association for Computer Linguistics, 2014.

[18] Tapomayukh Bhattacharjee, James M. Rehg, and Charles C. Kemp. Haptic classification and recognition of objects using a tactile sensing forearm. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4090–4097, 2012.

[19] Joydeep Biswas and Manuela M. Veloso. Localization and navigation of the cobots over long-term deployments. *The International Journal of Robotics Research*, 32(14):1679–1694, 2013.

[20] Michael Bowling and Manuela Veloso. Simultaneous adversarial multi-robot learning. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 699–704, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[21] S. R. K. Branavan, Harr Chen, Luke S. Zettlemoyer, and Regina Barzilay. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 82–90, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[22] Rodney A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1):139 – 159, 1991.

[23] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *CoRR*, abs/1805.11085, 2018.

[24] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 314–323. PMLR, 13–15 Nov 2017.

[25] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *In AAAI*, 2010.

[26] Ricson Cheng, Arpit Agarwal, and Katerina Fragkiadaki. Reinforcement learning of active vision for manipulating objects under occlusions. In *Conference on Robotic learning (CoRL)*, 2018.

[27] Wang Z. Cheng, R. and K. Fragkiadaki. Geometry-aware recurrent neural networks for active visual recognition. In *NIPS*, 2018.

[28] Samuel Clarke, Travers Rhodes, Christopher G. Atkeson, and Oliver Kroemer. Learning audio feedback for estimating amount and flow of granular material. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 529–550. PMLR, 29–31 Oct 2018.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[30] Siyuan Dong, Wenzhen Yuan, and Edward H Adelson. Improved gelsight tactile sensor for measuring geometry and slip. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 137–144. IEEE, 2017.

[31] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[32] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. *CoRR*, abs/1511.07404, 2015.

[33] Katerina Fragkiadaki, Jonathan Huang, Alex Alemi, Sudheendra Vijayanarasimhan, Susanna Ricco, and Rahul Sukthankar. Motion prediction under multimodality with conditional stochastic networks. *CoRR*, abs/1705.02082, 2017.

[34] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA, 1979.

[35] Alison Gopnik. The scientist in the crib: minds,brains, and how children learn. 1999.

[36] Alison Gopnik and Laura Schulz. Causal learning. 2007.

[37] Di Guo, Fuchun Sun, Bin Fang, Chao Yang, and Ning Xi. Robotic grasping using visual and tactile sensing. *Inf. Sci.*, 417(C):274–286, November 2017.

[38] David Ha and Jürgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018.

[39] Jessica Hamrick, Peter Battaglia, and Joshua B Tenenbaum. Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd annual conference of the cognitive science society*, pages 1545–1550. Cognitive Science Society Austin, TX, 2011.

[40] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[41] Mohamed K. Helwa and Angela P. Schoellig. Multi-robot transfer learning: A dynamical system perspective. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4702–4708, 2017.

[42] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[43] https://cloud.google.com/speech/.

[44] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2):21:1–21:35, April 2017.

[45] C. Kerl, J. Sturm, and D. Cremers. Dense visual SLAM for RGB-D cameras. In *IROS*, 2013.

[46] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[47] Kurt Koffka. *Principles of Gestalt Psychology*. New York: Harcourt, Brace, 1935.

[48] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[49] Jayant Krishnamurthy and Tom M. Mitchell. Joint syntactic and semantic parsing with combinatory categorial grammar. In *ACL*, 2014.

[50] O. Kroemer, CH. Lampert, and J. Peters. Learning dynamic tactile sensing with robust vision-based training. *IEEE Transactions on Robotics*, 27(3):545–557, June 2011.

[51] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *CoRR*, abs/1504.00702, 2015.

[52] Toby Jia-Jun Li, Igor Labutov, Brad A. Myers, and Tom M. Mitchell. Supporting co-adaptive human-agent relationship through programming by demonstration using existing guis (submitted). *Proceedings of the 2018 CHI Workshop on Rethinking Interaction*, 2018.

[53] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[54] Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 206–214. Curran Associates, Inc., 2012.

[55] Maja J. Matarić. Reinforcement learning in the multi-robot domain. *Auton. Robots*, 4(1):73–83, March 1997.

[56] Dipendra Kumar Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. *CoRR*, abs/1704.08795, 2017.

[57] Dipendra Kumar Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. In *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014.

[58] Dipendra Kumar Misra, Kejia Tao, Percy Liang, and Ashutosh Saxena. Environment-driven lexicon induction for high-level instructions. In *ACL*, 2015.

[59] T. Mitchell, W. Cohen, E. Hruscha, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohammad, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *AAAI*, 2015. : Never-Ending Learning in AAAI-2015.

[60] Tom M. Mitchell. Explanation based learning: A comparison of symbolic and neural network approaches. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 197–204. Morgan Kaufmann, 1993.

[61] Andrew W. Moore and Christopher G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13:103–130, 1993.

[62] J. Morimoto and C. G. Atkeson. Nonparametric representation of an approximated Poincaré map for learning biped locomotion. *Autonomous Robots*, 27(2):131–144, 2009.

[63] Chrystopher L. Nehaniv and Kerstin Dautenhahn. Imitation in animals and artifacts. chapter The Correspondence Problem, pages 41–61. MIT Press, Cambridge, MA, USA, 2002.

[64] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[65] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 488–489, 2017.

[66] Karime Pereida, Mohamed K. Helwa, and Angela P. Schoellig. Data-efficient multi-robot, multi-task transfer learning for trajectory tracking. *IEEE Robotics and Automation Letters*, 3(2):1260–1267, 2018.

[67] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *CoRR*, abs/1709.10087, 2017.

[68] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[69] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *CoRR*, abs/1606.04671, 2016.

[70] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. *arxiv*, 2018.

[71] S. Schaal. Is imitation learning the route to humanoid robots? 3(6):233–242, 1999.

[72] *S. Schaal and C. G. Atkeson. Learning control for robotics. *IEEE Robotics & Automation Magazine*, 17(2):20–29, 2010.

[73] S. Schaal, C. G. Atkeson, and S. Vijayakumar. Scalable locally weighted statistical techniques for real time robot learning. *Applied Intelligence*, 16(1), 2002.

[74] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *CoRR*, abs/0812.4360, 2008.

[75] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 802–810, Cambridge, MA, USA, 2015. MIT Press.

[76] Maximilian Sieb and Katerina Fragkiadaki. Data dreaming for object detection: Learning object-centric state representations for visual imitation. In *Humanoids*, 2018.

[77] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.

[78] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[79] Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Universal planning networks. *CoRR*, abs/1804.00645, 2018.

[80] Shashank Srivastava, Igor Labutov, and Tom M. Mitchell. Joint concept learning and semantic parsing from natural language explanations. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[81] Bradly C. Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *CoRR*, abs/1703.01703, 2017.

[82] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, 2012.

[83] Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '11, pages 761–768, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems.

[84] Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning spatial common sense with geometry-aware recurrent neural networks. In *arxiv*, 2018.

[85] Fish Tung and Katerina Fragkiadaki. Reward learning using natural language. *CVPR*, 2018.

[86] Manuela Veloso, Joydeep Biswas, Brian Coltin, and Stephanie Rosenthal. CoBots: Robust Symbiotic Autonomous Mobile Service Robots. In *Proceedings of IJCAI'15, the International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, July 2015.

[87] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson. 3d shape perception from monocular vision, touch, and shape priors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1606–1613, Oct 2018.

[88] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 153–164. Curran Associates, Inc., 2017.

[89] A. Yamaguchi and C. G. Atkeson. Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables. In *IEEE-RAS International Conference on Humanoid Robotics*, 2016.

[90] *A. Yamaguchi and C. G. Atkeson. Differential dynamic programming for graph-structured dynamical systems: Generalization of pouring behavior with different skills. In *IEEE-RAS International Conference on Humanoid Robotics*, 2016.

[91] *A. Yamaguchi and C. G. Atkeson. Model-based reinforcement learning with neural networks on hierarchical dynamic system. In *the Workshop on Deep Reinforcement Learning: Frontiers and Challenges in the 25th International Joint Conference on Artificial Intelligence (IJCAI2016)*, 2016.

[92] *A. Yamaguchi and C. G. Atkeson. Neural networks and differential dynamic programming for reinforcement learning problems. In *the IEEE International Conference on Robotics and Automation (ICRA'16)*, 2016.

[93] Akihiko Yamaguchi and Christopher G. Atkeson. Implementing tactile behaviors using fingervision. In *IEEE-RAS International Conference on Humanoid Robotics*, 2017.

[94] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[95] Wenzhen Yuan, Rui Li, Mandayam A Srinivasan, and Edward H Adelson. Measurement of shear and slip with a gelsight tactile sensor. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 304–311. IEEE, 2015.

[96] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[97] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR17), Honolulu, HI, USA*, pages 21–26, 2017.

[98] Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A Srinivasan, and Edward H Adelson. Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 951–958. IEEE, 2017.

[99] Dongqing Zhang, Ilknur Icke, Belma Dogdas, Sarayu Parimal, Smita Sampath, Joseph Forbes, Ansuman Bagchi, Chih-Liang Chin, and Antong Chen. A multi-level convolutional LSTM model for the segmentation of left ventricle myocardium in infarcted porcine cine MR images. *CoRR*, abs/1811.06051, 2018.

[100] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. *CoRR*, abs/1710.04615, 2017.

KATERINA FRAGKIADAKI
Assistant Professor
Machine Learning Department, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, 15213
2675289476
katef@cs.cmu.edu, www.cs.cmu.edu/∼katef

Professional Preparation

National Technical University of Athens	EECS	**Diplomat** 2007
University of Pennsylvania	CIS	**M.S.** 2011
University of Pennsylvania	CIS	**Ph.D.** 2013
EECS, UC Berkeley	PostDoctoral Fellow,	2013–2015
Google Research	PostDoctoral Fellow,	Oct. 2015–December 2016

Appointments

Assistant Professor	MLD, CMU	September 2016–present

Products Five Closely Related products to the proposed project

1. Tung H.-Y. F., Cheng R., Fragkiadaki K., 2018, Learning Spatial-Common Sense with Geometry-Aware Recurrent Neural Networks, *in submission*
2. Cheng R., Wang Z., Fragkiadaki K., 2018, Geometry-Aware Recurrent Neural Networks for Active Visual Recognition , *Neural Information Processing Systems (NIPS)*
3. Tung H.-Y. F., Harley A. Seto W., Fragkiadaki K., 2017, Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision, *International Conference on Computer Vision (ICCV)*
4. Cheng R., Agarwal A., Fragkiadaki K., 2018, Reinforcement Learning of Active Vision for Manipulating Objects under Occlusions, *Conference on Robotic Learning (CoRL)*
5. Fragkiadaki K., Agrawal P., Levine S., Malik J., 2016, Learning Visual Predictive Models of Physics for Playing Billiards, *International Conference to Learning Representations(ICLR)*

Five Other Products

1. Fragkiadaki K., Levine S., Felsen P., Malik J., 2015, Recurrent Network Models for Human Dynamics, *IEEE International Conference on Computer Vision (ICCV)*
2. Vijayanarasimhan S., Ricco S., Schmid C., Sukthankar R., Fragkiadaki K., 2017, SfM-Net: Learning of Structure and Motion from Video, *arxiv*
3. Fragkiadaki K., Salas M., Arbelaez P., Malik J., 2014, Grouping-based Low-Rank Trajectory Completion and 3D Reconstruction, *Neural Information Processing Systems (NIPS)*
4. Tung H.-Y. F., Tung W., Yumer E., Fragkiadaki K., 2017, Self-supervised learning of Motion Capture, *Neural Information Processing Systems (NIPS)*
5. Fragkiadaki K., Arbelaez P., Felsen P., Malik J., 2015, Learning to Segment Moving Objects in Videos, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

Synergistic Activities-Awards

- Google Faculty Award 2018, UPMC faculty Award 2019

- Organizer of CMU AI4ALL summer school 2018/2019
- Area Chair for CVPR 2018, ICLR 2019, ICML 2019

- Organizer: The 11th Perceptual Organization for Computer Vision Workshop, CVPR 2016: ”The role of feedback in Recognition and Segmentation”. Workshop that brought together Human and Computer Vision scientists to investigate incorporation of Feedback in visual architectures
- Best Ph.D. Thesis, Computer and Information Science Department, University of Pennsylvania, 2013.

CHRISTOPHER GRANGER ATKESON

Professor
The Robotics Institute and the Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, 15213
cga@cmu.edu, www.cs.cmu.edu/∼cga

Education

Harvard	Biochemistry	**A.B.** 1981, summa cum laude
Harvard	Applied Mathematics (Computer Science)	**S.M.** 1981
MIT	Brain and Cognitive Sciences	**Ph.D.** 1986

Employment

Professor	RI & HCII, CMU	2004–present
Associate Professor	RI & HCII, CMU	2000–2004
Associate Professor	College of Computing, Georgia Tech	1994–2000
Associate Professor	Brain and Cognitive Sciences, MIT	1990–1993
Assistant Professor	Brain and Cognitive Sciences, MIT	1986–1990

Five Related Publications

1. "Implementing tactile behaviors using FingerVision", A. Yamaguchi and C. G. Atkeson, *IEEE-RAS 17th International Conference on Humanoid Robots (Humanoids),* 2017.
2. "Design of a Lightweight Soft Robotic Arm Using Pneumatic Artificial Muscles and Inflatable Sleeves", P Ohta, L Valle, J King, K Low, J Yi, C G Atkeson, and Y-L Park, *Soft Robotics,* 5(2): 204-215, 2018.
3. "Team WPICMU: Achieving Reliable Humanoid Behavior in the DARPA Robotics Challenge", M. DeDonato, F. Polido, K. Knoedler, B.P.W. Babu, N. Banerjee, C.P. Bove, X. Cui, R. Du, P. Franklin, J.P. Graff, P. He, A. Jaeger, L. Li, D. Berenson, M.A. Gennert, S. Feng, C. Liu, X. Xinjilefu, J. Kim, C.G. Atkeson, X. Long, and T. Padir, *Journal of Field Robotics* 34 (2), 381-399, 2017.
4. "Using Deep Reinforcement Learning to Learn High-Level Policies on the ATRIAS Biped", T. Li, A. Rai, H. Geyer, and C. G. Atkeson, arXiv preprint arXiv:1809.10811, 2018.
5. "Interpersonal interactions for haptic guidance during balance exercises", S.M. Steinl, P.J. Sparto, C.G. Atkeson, M.S. Redfern, and L. Johannsen, *Gait & Posture* 65, 129-136, 2018.

Five Other Relevant Publications

1. "Human-in-the-loop optimization of exoskeleton assistance during walking", J. Zhang, P. Fiers, K. A. Witte, R. W. Jackson, K. L. Poggensee, C. G. Atkeson, and S. H. Collins, *Science,* 356:6344, 1280–1284, 2017.
2. "Learning Tasks From Observation and Practice", D. C. Bentivegna, C. G. Atkeson, and G. Cheng, *Robotics and Autonomous Systems,* 47:163-169, 2004.
3. "Finding and Transferring Policies Using Stored Behaviors", M. Stolle and C. G. Atkeson, *Autonomous Robots,* 29(2): 169-200, 2010.

4. "Learning Control for Robotics", S. Schaal and C. G. Atkeson, *IEEE Robotics & Automation Magazine,* 17(2), 20-29, 2010
5. "Efficient Robust Policy Optimization", C. G. Atkeson, *American Control Conference (ACC),* 2012.

Synergistic Activities

- *IEEE-RAS International Conference on Humanoid Robots:* Program Committee Co-Chair, 2003, General Chair, 2004, US Program Committee Chair, 2008, General Co-Chair, 2012.
- Scientific Board, *Dynamic Walking Conference,* 2005-present.

Tom M. Mitchell
E. Fredkin University Professor
Machine Learning Department
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
Telephone: (412) 2682611, (412) 268-1299, (412) 721-2414
Email: tom.mitchell@cmu.edu

Professional Preparation:

Massachusetts Institute of Technology	Electrical Engineering	S.B., 1973
Stanford University	Electrical Engineering	M.S., 1975
Stanford University	Electrical Engineering, Computer Science minor	Ph.D., 1979

Appointments:

2009-present	E. Fredkin University Professor, Carnegie Mellon University
2006-2016	Department Head, Machine Learning Department, Carnegie Mellon University
1999-2008	E. Fredkin Professor of Machine Learning, Carnegie Mellon University
19972006	Director, Ctr. for Automated Learning and Discovery, Carnegie Mellon University
19861999	Professor, Computer Science and Robotics, CarnegieMellon University
2000-2002	Vice President and Chief Scientist, WhizBang! Labs, Pittsburgh, PA.
197886	Assistant/Associate Professor, Department of Computer Science, Rutgers University,

Five Related Products:

- *Instructable Intelligent Personal Agent.* Amos Azaria, Jayant Krishnamurthy, and Tom M. Mitchell. Proceedings of the AAAI Conference, 2016.
- *Joint concept learning and semantic parsing from natural language explanations*. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017
- *An end user development approach for failure handling in goal-oriented conversational agents*. Toby Jia-Jun Li, Igor Labutov, Brad A. Myers, Amos Azaria, Alexander I. Rudnicky, and Tom M. Mitchell. In Robert J. Moore, Margaret H. Szymanski, Raphael Arar, and Guang-Jie Ren, editors, ***Studies in Conversational UX Design***. Springer (in press), 2018.
- *Machine Learning: Trends, Perspectives, and Prospects.* M. I. Jordan and T. M. Mitchell, ***Science*** 349:255, July 2015. DOI: 10.1126/science.aaa8415.
- *Machine Learning*, T.M. Mitchell, McGraw Hill, 1997. (textbook)

Service, Synergistic Activities:

- Currently lead a research project on sensor-effector agents learning from user verbal instruction, which is exploring how to enable users of mobile phones to teach their phones new capabilities. This project involves teaching a "softbot" instead of a "robot" but shares many fundamental research issues.
- Serve as Co-PI (along with Professor Justine Cassell) on the $10M InMind project which seeks to develop prototypes of future intelligent mobile device agents. This project involves over a dozen diverse faculty-led projects that bear on user-agent interactions, teaching and customization, which lies at the core of the research we propose here.

- Founded CMU’s Machine Learning Department, which offers the world’s first Ph.D. program in Machine Learning, increasing both the pool of researchers available to support the proposed research, and its multidisciplinary educational impacts.
- PI of the Never Ending Language Learning (NELL) research project, an effort to develop a computer to learn continuously for years to read the web. NELL has been running 24 hours/day since 2010, and has produced a collection of over 120 million beliefs.
- Co-Chaired recent U.S. National Academy study on “Information Technology, Automation, and the U.S. Workforce” and testified to the U.S. Congressional Research Service on results of the study.

Wenzhen Yuan

Assistant Professor
The Robotics Institute
5000 Forbes Avenue, Pittsburgh PA 15213-3890

yuanwz@cmu.edu
https://people.csail.mit.edu/yuan_wz

EDUCATION

Tsinghua University	Beijing, China	Mechanical Engineering	B.Eng. 2012
MIT	Cambridge, MA	Mechanical Engineering	S.M. 2014
MIT	Cambridge, MA	Mechanical Engineering	Ph.D. 2018

APPOINTMENTS

Aug. 2019-	Assistant Professor, The Robotics Institute, Carnegie Mellon University
Oct. 2018-Jul. 2019	Postdoctoral Scholar, Computer Science Department, Stanford University

PRODUCTS

Five Relevant Publications

- Yuan, W., Dong, S., and Adelson, E. H. (2017). Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors, 17*(12), 2762.
- Yuan, W., Wang, S., Dong, S., and Adelson, E. H. (2017, July). Connecting Look and Feel: Associating the visual and tactile properties of physical materials. In *CVPR* (pp. 4494-4502).
- Yuan, W., Zhu, C., Owens, A., Srinivasan, M. A., and Adelson, E. H. (2017, May). Shape-independent hardness estimation using deep learning and a GelSight tactile sensor. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on* (pp. 951-958).
- Wang, S., Wu, J., Sun, X., Yuan, W., Freeman, W. T., Tenenbaum, J. B., and Adelson, E. H. (2018, October). 3D shape perception from mnocular vision, touch, and shape priors. In *Intelligent Robots and Systems (IROS 2018), 2018 IEEE/RSJ International Conference on*
- Yuan, W., Li, R., Srinivasan, M. A., and Adelson, E. H. (2015, May). Measurement of shear and slip with a GelSight tactile sensor. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on* (pp. 304-311).

Five Other Significant Publications

- Calandra, R., Owens, A., Upadhyaya, M., Yuan, W., Lin, J., Adelson, E. H., and Levine, S. (2017, October). The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes?. In *Conference on Robot Learning* (pp. 314-323).
- Yuan, W., Mo, Y., Wang, S., and Adelson, E. H. (2018, May). Active Clothing Material Perception using Tactile Sensing and Deep Learning. In *Robotics and Automation (ICRA), 2018 IEEE International Conference on* (pp. 1-8).

- Luo, S., Yuan, W., and Adelson, E. H., Cohn, A. G., and Fuentes, R. (2018, May). ViTac: Feature Sharing Between Vision and Tactile Sensing for Cloth Texture Recognition. In *Robotics and Automation (ICRA), 2018 IEEE International Conference on* (pp. 2722-2727).

- Dong, S., Yuan, W., and Adelson, E. H. (2017, September). Improved gelsight tactile sensor for measuring geometry and slip. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on* (pp. 137-144).

- Li, R., Platt, R., Yuan, W., ten Pas, A., Roscup, N., Srinivasan, M. A., and Adelson, E. (2014, September). Localization and manipulation of small parts using gelsight tactile sensing. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on* (pp. 3988-3993).

SYNERGISTIC ACTIVITIES

- Co-organizer of workshop Tactile Sensing for Manipulation: Hardware, Modeling and Learning at the conference of Robotics: Science and Systems (RSS), 2017, Cambridge, MA.
- Actively promoting the research and application in robotic tactile sensing. Founded the open-source platform for GelSight tactile sensor and shared the hardware platform with researchers from 9 universities or research institutes. The support and collaboration resulted in 5 published papers.
- Served as peer reviewer of 3 high-impact robotics or sensor related journals: Mechatronics, Sensors & Actuators: A: Physical, IEEE Robotics and Automation Letters (RA-L) and 3 top robotics conferences: International Conference on Intelligent Robots and Systems (IROS), IEEE-RAS International Conference on Humanoid Robots (Humanoids),and Conference on Automation Science and Engineering (CASE).
- Contributor of the C++ robotics toolbox Drake. Drake is an open source toolbox for analyzing the dynamics of robots and building control systems for them. Now it is used by multiple research groups across US.

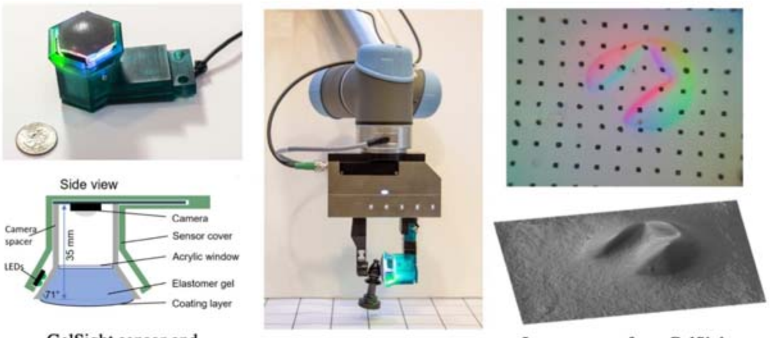
CMU Facilities

We will use three existing testbeds to evaluate the proposed learning ecosystem. We describe these testbeds and their corresponding robots in this section. The lab-based **Deformable/Liquid/Granular Testbed** involves existing human-scale robots working with deformable, liquid, and granular materials as is found in food preparation and science experiment kits for children. These robots, grippers, and tactile sensors include Baxters, Sawyers, Franka Emika Pandas, Robotiq hands, parallel jaw grippers, and FingerVision and GelSight (Figure 1) tactile sensors. We use extensive robot-mounted and external cameras, which at this point are too cheap and plentiful to mention. The **Social/Affordance Testbed** is based on our existing deployment of CoBots in our computer science building. We have developed low cost robots for our **Toys/Kits Testbed.** We describe our existing effort on low cost robots for research and education, and the educational toys and kits they are working with. We describe our behavior capture facilities including the Motion Capture Lab (based on reflective markers and IR illumination) and the Panoptic Studio (based on several hundred video cameras), which are used to capture human teacher behavior and human-robot interaction. We also describe our work on virtual reality-based robot training.

1 The Deformable/Liquid/Granular Testbed and lab-based robots

In this testbed human-scale robots work with deformable, liquid, and granular materials as are found in food preparation and science experiment kits for children. We will use a number of Baxter, Sawyer, and Franka Emika Panda robots which we have access to for this study. We also have several robot hands, such as Robotiq hands and parallel jaw grippers. Atkeson has developed FingerVision tactile sensors, and Yuan has GelSight sensors (Figure 1) and plans to develop next-generation vision-based tactile sensors. In a separate project we are building new hands which may be useful to this project (see **Results from Prior NSF Support**).

The physical testbeds we will use for evaluation will each include at least two robot arms with hands. An example setup is a Baxter research robot with FingerVision tactile sensors mounted on its fingers (Figure 2). It has two 7 DOF arms, and two different types of parallel jaw grippers. Its arm payload is 2.2 kg. One gripper's grip force is 44 N with a grip range 37 to 75 mm, and the other



GelSight sensor and design scheme

GelSight on a robot

Image output from GelSight, and reconstructed 3D shape

Figure 1: The GelSight tactile sensor which measures high-resolution geometry of the contact surface from the image-formatted output. The black dots painted on the sensor surface is used to estimate the contact force.

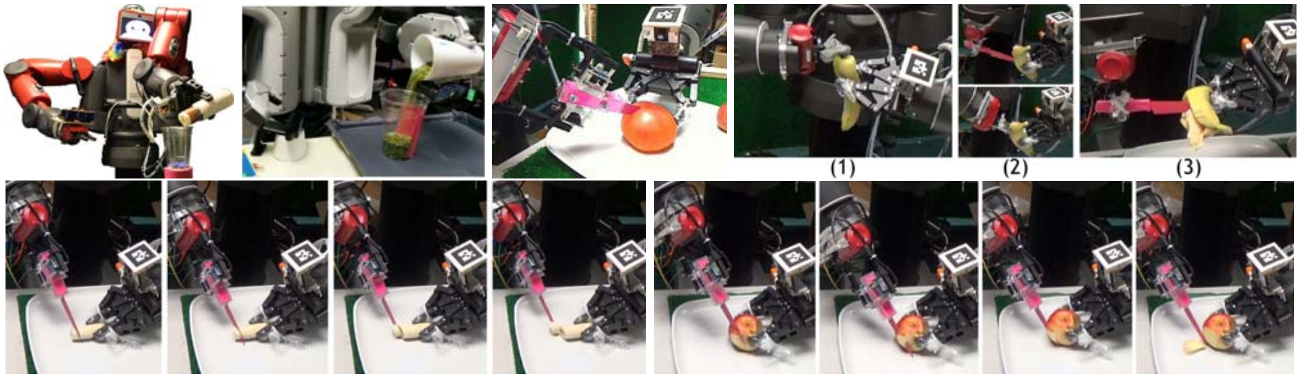


Figure 2: One of our Baxter robots pouring, our PR2 robot pouring, cutting a tomato, trying to peel a banana using teleoperation (1 is a human-like strategy of ripping the skin at the stem, and 2 and 3 explore alternative strategies more appropriate for the robot), cutting a peeled banana (easy), and cutting an apple (hard).

gripper's grip force is 100 N with a grip range range 0 to 84 mm. The Baxter robot can estimate torque at each joint. Here is a video of our Baxter robot pouring: `https://youtu.be/NIn-mCZ-h_g` We have instrumented Baxter's fingers with *FingerVision,* our optically based tactile sensing and proximity vision system. This video of the system is a good introduction to FingerVision: `https://youtu.be/ifOwQdy9gDg` The vision system both measures the deformation of skin on the robot fingers to measure contact locations and forces, and sees through the transparent skin to provide object and surface localization and tracking during manipulation (in this case cutting with a knife). We also have a number of soft robot prototypes currently available for experimentation and others under development.

The Humanoid Robotics Lab (located in the Motion Capture Lab described below) provides a state of the art full-sized humanoid for research and education. We have developed a hydraulic humanoid in collaboration with Sarcos (with NSF equipment funding). We use this robot because of the speed, power, and achievable joint compliance of the hydraulics and the range of motion of the joints.

We will deliberately test our ideas using tasks and materials that are hard to model, such as deformable, liquid, and granular materials (Figure 3). We believe cultural knowledge (knowledge learned from others) of how to do tasks is more important than planning or reasoning from scratch for these materials. Our previous work has focused on manipulating liquids and particulate materials (for example, pouring), and cutting objects with skins. These tasks also allowed us to explore learning different task strategies. In the case of pouring, strategies include tipping, shaking, and tapping. In the case of cutting, one can slice, stab, or saw. We believe humans learn these type of skills from demonstrations by other humans, and we all maintain skill libraries representing alternative strategies to do the same task.

For this testbed we have chosen food preparation and science experiment kits as the tasks (Figure 3). We have chosen food preparation as a good domain, partly because we are already working on food preparation in collaboration with Sony. We have chosen making salads as an initial focus task because there is an existing infrastructure and progression of difficulty: one can start with "kit/bag/box salads" where all items necessary can be purchased pre-washed and pre-cut in a bag or box, progress to working with bulk pre-cut materials, and graduate to preparing a salad from whole vegetables. Similarly, with baking we can progress from baking kits for children to cookie, cake, and muffin mixes commonly found in supermarkets, to preparing materials "from scratch". Science experiment kits for

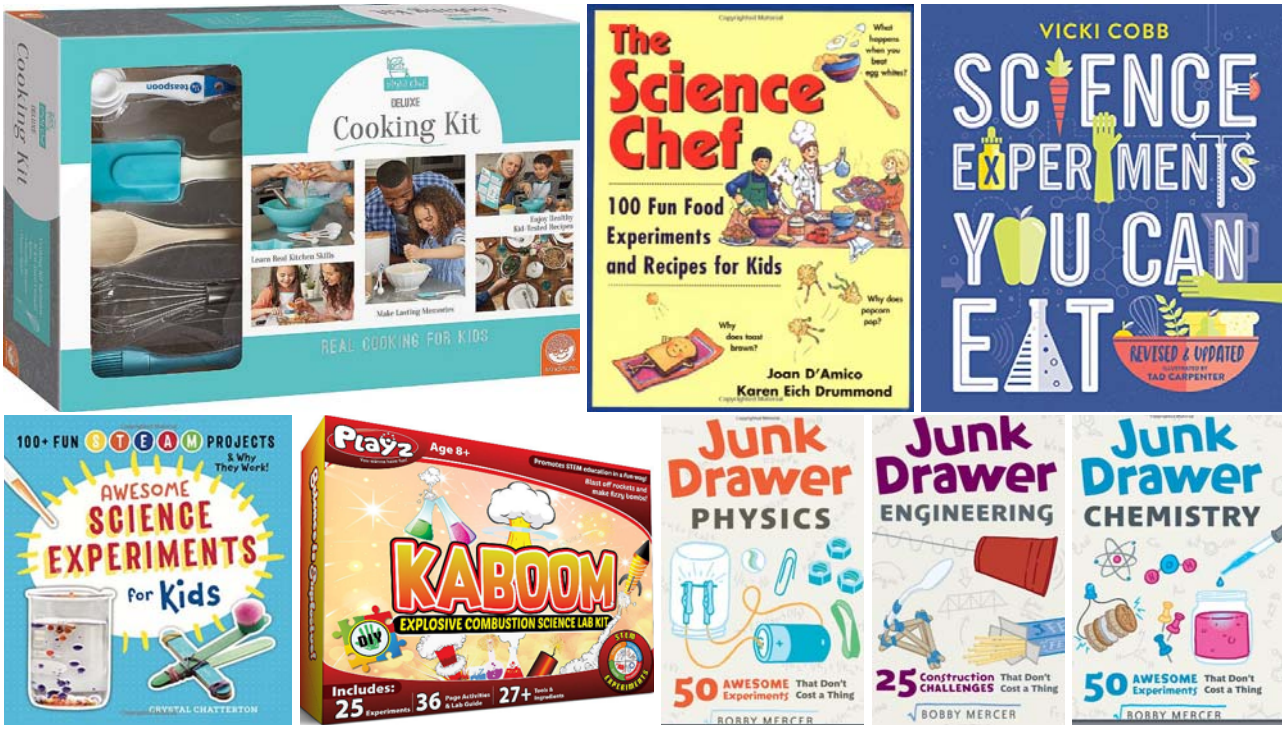

DELUXE
Cooking Kit
REAL COOKING FOR KIDS
The Science Chef
100 Fun Food Experiments and Recipes for Kids
Joan D'Amico
Karen Eich Drummond
VICKI COBB
SCIENCE EXPERIMENTS YOU CAN EAT
REVISED & UPDATED
100+ FUN STEAM PROJECTS & Why They Work!
AWESOME SCIENCE EXPERIMENTS for Kids
CRYSTAL CHATTERTON
Playz
Age 8+
KABOOM
EXPLOSIVE COMBUSTION SCIENCE LAB KIT
Includes: 25 Experiments 36 27+
Junk Drawer PHYSICS
50 AWESOME Experiments That Don't Cost a Thing
BOBBY MERCER
Junk Drawer ENGINEERING
25 Construction CHALLENGES That Don't Cost a Thing
BOBBY MERCER
Junk Drawer CHEMISTRY
50 AWESOME Experiments That Don't Cost a Thing
BOBBY MERCER

Figure 3: Some domains we will use to inspire and evaluate our work, taken from educational kits and books involving physical processes and devices a robot can work with, learn and reason about, and repair, including mechanical, electrical, thermal, chemical, and combustion processes.

children allow robots to explore mechanics, chemistry, and electricity while performing manipulations expected of young children. Food preparation and science experiments will also be learned from online recipes, cooking videos, and instructional materials, which opens up a huge range of instructional material.

2 The Social/Affordance Testbed and wandering CoBots

We currently have multiple CoBot service robots (cobots, see Figure 4 and 5d) capable of performing user-requested tasks in our multi-floor office buildings. These mature platforms are useful for the proposed work because they provide a reliable platform on which we can focus on the interaction between the users and the robots. One of the distinguishing features of the CoBot platform is the stable low-clearance and omnidirectional base which makes it well-suited for its role as a guide. The base is a scaled-up version of the CMDragons small-size soccer robot. The robots have operated for more than 1,000km and for more than three years without hardware failures, and with minimal maintenance.

The CoBots can already perform multiple classes of tasks which involve social interaction and knowing object affordances. Tasks are programmed by humans through a speech or web interface. Cobots currently ask for help from humans for any manipulation or object detection task, as they are not equipped with manipulation hardware or software, or semantic perceptual capabilities. Tasks include: 1) A single destination task, in which the user asks the robot to go to a specific location the Go-To-Room task and, in addition, to deliver a specified spoken message the Deliver-Message task; 2) A

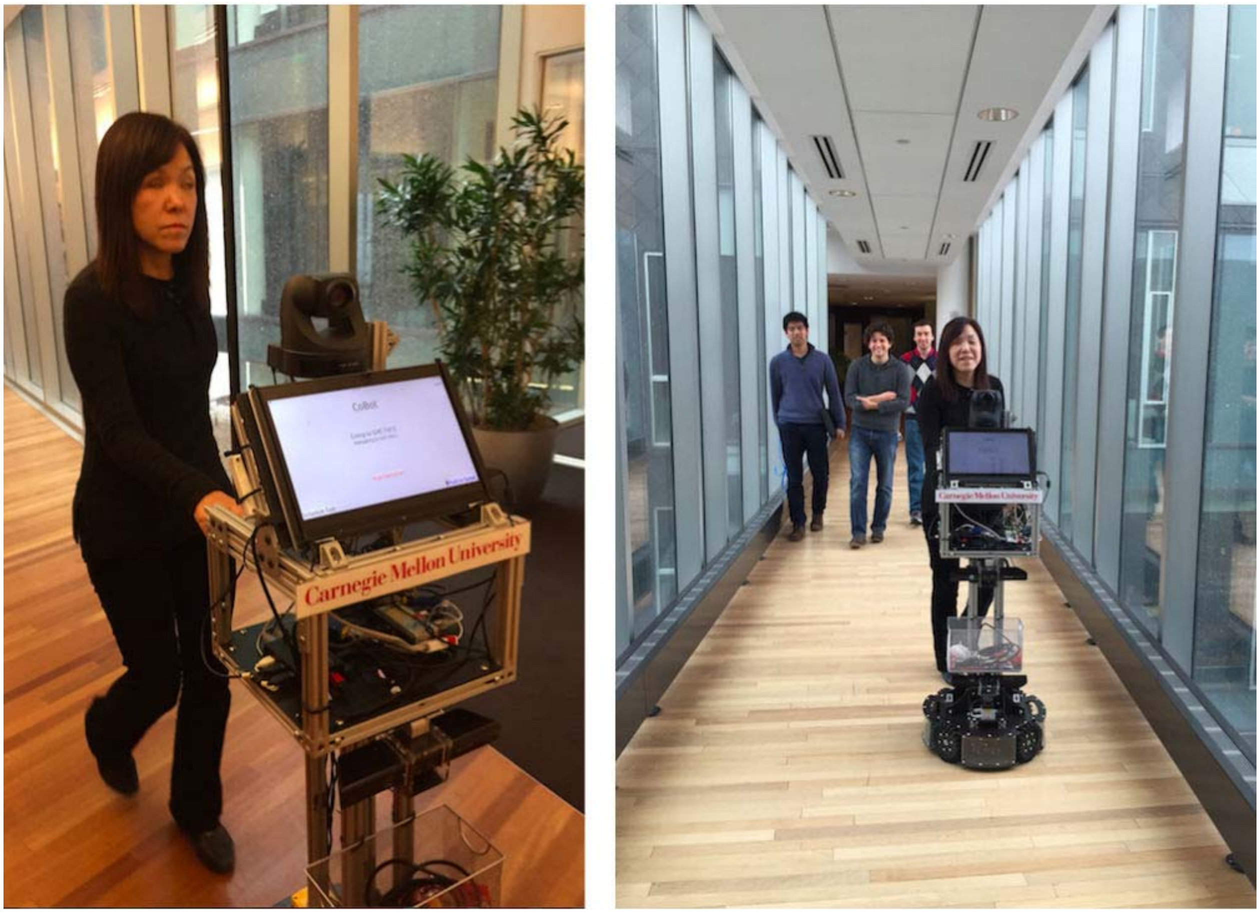


Figure 4: A CoBot leading a blind user.

Transport task, in which the user requests the robot to retrieve an item and to deliver it to a destination location. This Transport task is also used for accompanying a person between locations (when the item to transport is a person). A task to escort a person to a specified location, the Escort task, in which the robot waits for a person in front of the elevator on the floor of the destination location, and guides the person to the location. Another task is the semi-autonomous Telepresence task, in which users may request to be remotely present on the mobile robot with autonomous navigation and obstacle avoidance. Users select destination points on the map or on the robot image view to move remotely through the telepresence web interface. Furthermore, they can control the robot through a rich motion- and perception-controlled web-based interface. The research underlying the autonomous CoBot robots has been focused on achieving a complete robust localization and navigation, so that the robots can move in our environments completely autonomously. The robots can detect obstacles and peoples silhouettes in particular. The research has strong underlying assumptions that surrounding humans can actually see the robots. Our previous work has demonstrated instructable navigation capabilities.

Planned CoBot upgrades: Using other funding sources, we will equip each CoBot with a Kinova Jaco arm, a three finger gripper, and over-the-shoulder, wrist, and palm cameras, in addition to the current body-mounted Kinect camera. Our goal is to teach the CoBots a diverse set of navigation and manipulation tasks, natural language understanding and much enhanced vision capabilities, as well as the ability to continually adapt their behaviors to the preferences of the users. Success of the proposed research will be evaluated on the basis of the CoBots' acquired competence in language,

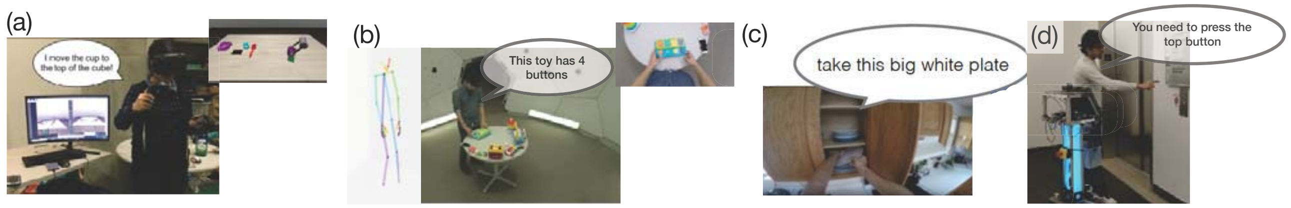

(a)
I move the cup to the top of the cube!
(b)
This toy has 4 buttons
(c)
take this big white plate
(d)
You need to press the top button

Figure 5: Collecting narrated demonstrations a) in virtual reality, b) in an instrumented space (the CMU Panoptic Studio), c) in users' homes captured by egocentric mounted cameras, and d) interactively with a CoBot.

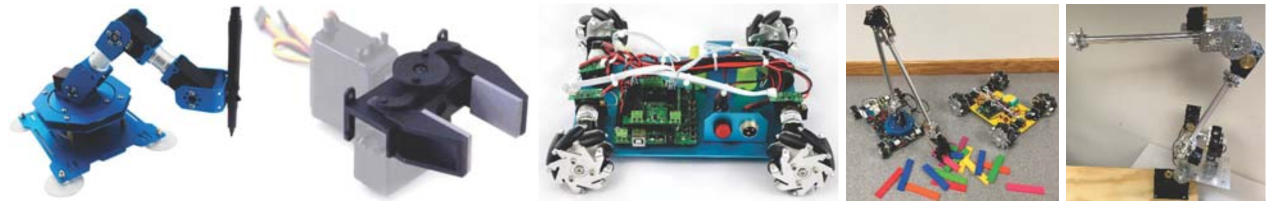


Figure 6: **Low cost robot prototypes:** see text.

visual recognition and behavior learning while minimizing human teacher supervision over time, as well as the quality of their interaction with the human collaborators over time.

Engaging our community in CoBot teaching: Aside from the technological innovations, an additional aim of this research is to informally explore the development of a "humans teaching robots" culture. Will the users in our building be willing to teach the robots? Will they learn to be better teachers? The proposed research includes an informal social "experiment" for large scale human-robot collaborative learning. To reinforce engagement of our community, we will set up "robot museum" exhibits that explain how the robots work and provide insight into what they are learning from the humans.

3 The Toys/Kits Testbed and low cost robots

We are already building a variety of low cost robots (1000-2000$), which are being used in a project class *16-264 Humanoids* taught by Atkeson (Figure 6). The left 3 photos show the commercial robot parts used as ingredients: a) a LewanSoul Drawarm ($140, we can also use the LewanSoul Xarm, $200), b) a Lynxmotion "Little Grip" gripper ($16), and c) a Nexusrobot 4WD 60mm Mecanum Wheel Arduino Robot ($330). A prototype showing the combination of these items is shown in the fourth photo, as well as a larger and beefier Nexusrobot ($630). The current reach of the modified Drawarm/Xarm is a little less than a meter (0.91m) and is easily adjustable. The last photo shows our prototype of a more heavily geared arm (stronger but slower than the modified Drawarm/Xarm). These arms have similar performance to any low cost arm built from hobby servos: they are slow, they are not accurate, it is difficult to improve the joint-level control, and they suffer from structural vibration which can be reduced by adding structural damping. However, these robots can perform the tasks required for many infant and K-12 educational toys and kits, especially with visual, tactile, and auditory feedback. We will use low cost handheld "tools" for our robots to extend their capabilities. Examples include solenoids for snapping together Lego parts and for "shooting" balls in small-scale versions of billiards, croquet, and mini (put-put) golf.

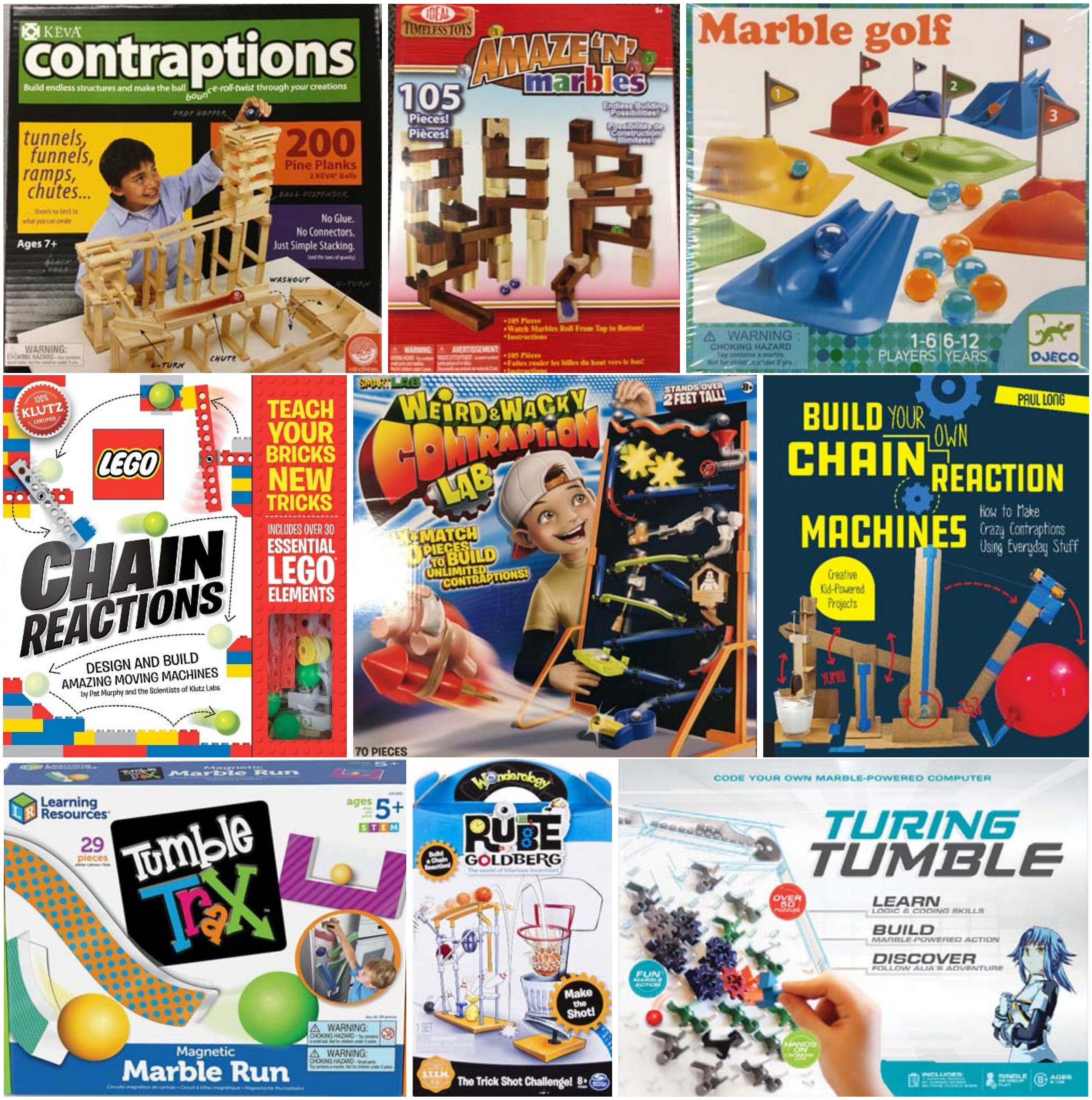
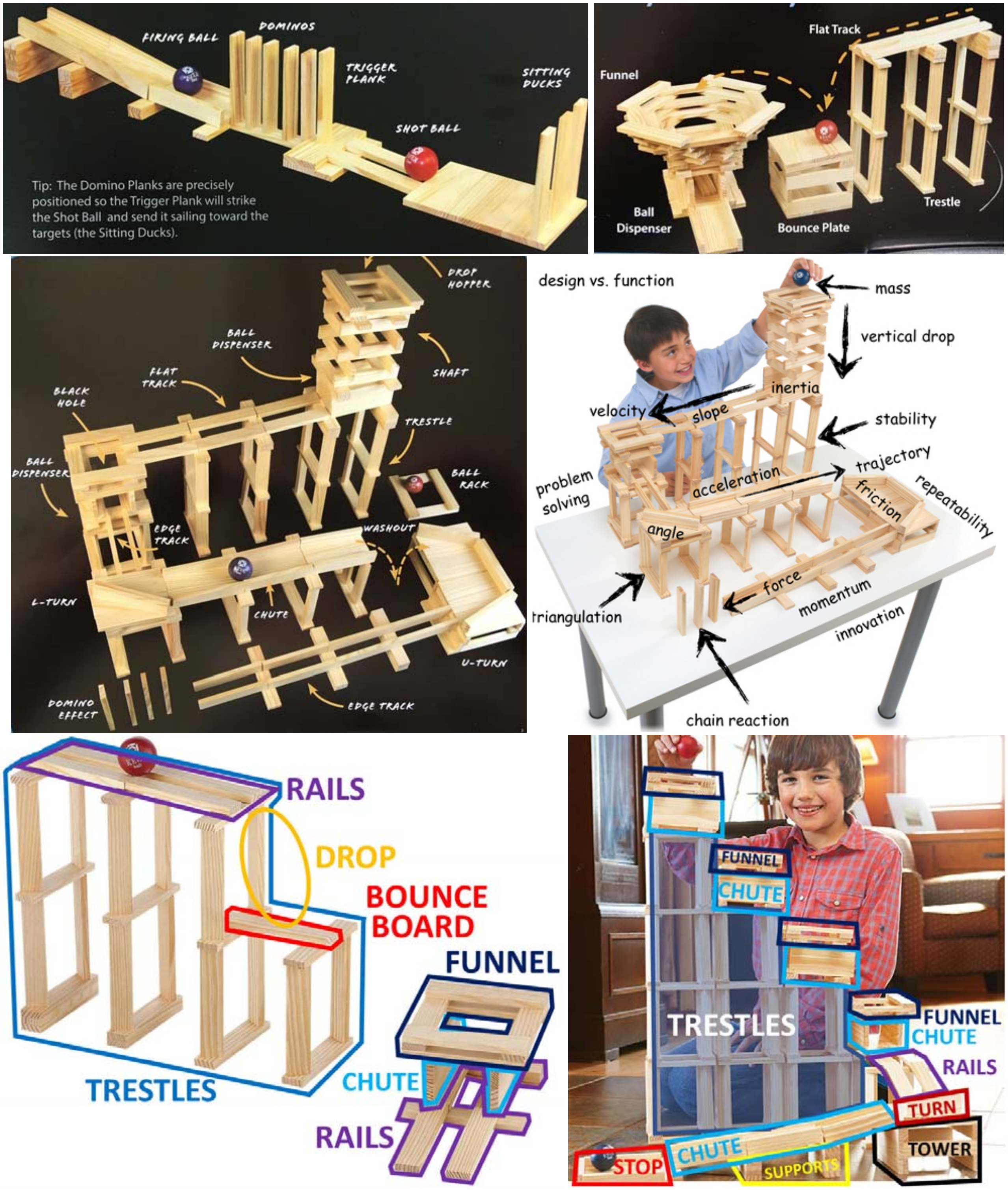

KEVA
contraptions
Build endless structures and make the ball bounce-roll-twist through your creations
tunnels, funnels, ramps, chutes...
200 Pine Planks
2 KEVA Balls
No Glue. No Connectors. Just Simple Stacking.
Ages 7+
WARNING:
IDEAL TIMELESS TOYS
AMAZE 'N' marbles
105 Pieces! Pièces!
Endless Building Possibilities!
Marble golf
1-6 PLAYERS
6-12 YEARS
DJECO
WARNING: CHOKING HAZARD
KLUTZ
LEGO
CHAIN REACTIONS
DESIGN AND BUILD AMAZING MOVING MACHINES
by Pat Murphy and the Scientists of Klutz Labs
TEACH YOUR BRICKS NEW TRICKS
INCLUDES OVER 30 ESSENTIAL LEGO ELEMENTS
SMART LAB
WEIRD & WACKY CONTRAPTION LAB
STANDS OVER 2 FEET TALL!
8+
MIX & MATCH 70 PIECES TO BUILD UNLIMITED CONTRAPTIONS!
70 PIECES
PAUL LONG
BUILD YOUR OWN CHAIN REACTION MACHINES
How to Make Crazy Contraptions Using Everyday Stuff
Creative Kid-Powered Projects
Learning Resources
29 pieces
ages 5+
STEM
Tumble Trax
Magnetic Marble Run
WARNING: CHOKING HAZARD
Wonderology
RUBE GOLDBERG
Make the Shot!
The Trick Shot Challenge!
8+
CODE YOUR OWN MARBLE-POWERED COMPUTER
TURING TUMBLE
LEARN LOGIC & CODING SKILLS
BUILD MARBLE-POWERED ACTION
DISCOVER FOLLOW ALIA'S ADVENTURE
OVER 50 PUZZLES
FUN MARBLE ACTION
HANDS ON SCREEN FREE

Figure 7: K-12 kits we are using.



Tip: The Domino Planks are precisely positioned so the Trigger Plank will strike the Shot Ball and send it sailing toward the targets (the Sitting Ducks).

Figure 8: **Top two rows:** Examples of labelling in instructional material in Keva Contraptions kits. **Bottom row:** Examples of desired visual segmentation into re-usable parts.

We will use a variety of educational toys and kits about mechanical processes and devices to develop and evaluate our ideas and algorithms about how robots can learn about perception, affordances, models, and skills. There are several reasons for this choice of task. These rigid body toys and kits are easy to simulate and replicate in virtual reality. They are widely available in a standardized form. The chosen tasks can be done by low cost (and low performance) robots. There is substantial instructional material on the web, including Youtube videos. Educational toys and kits are designed to stimulate and facilitate perceptual, affordance, model, skill, and cognitive learning. Toys for infants typically demonstrate numerous manipulation skills and affordances (knobs to turn, buttons to press, ...), are safe to operate, can break and be replaced easily and inexpensively, and use exaggerated visual and audio cues to guide the learner. Educational K-12 toys and kits we will focus on include toys and kits that focus on mechanical processes and construction such as Keva "Contraptions", Marble runs, Lego "Chain Reactions", Rube Goldberg kits, and Jenga (Figure 7). **Joke:** We will be focusing on toy problems and blocks world domains. In this case this is a good thing.

It is important to note that working with these kits will not require fast movement from the robot. The robot can move as slowly and carefully as necessary to successfully set up and initiate the physical process. After that, the robot mostly observes what happens and then hunts down stray marbles and other pieces (yes, a robot can lose its marbles). We will also focus on learning to play simple musical instruments such as keyboards, drums, rattles, and xylophones to emphasize multimodal visual, aural, and tactile learning. We will extend these domains with the robot's ability to design and 3D-print new components and tools as part of the learning process.

For all of these domains there is a rich set of instructional material and videos, as well as suggested experiments and other activities that can be used to inspire exploration, play, and evaluation. Figure 8 shows the kinds of linguistic labels found in the instructional material (top 2 rows), as well as manual labelling of re-usable parts (bottom row, we hope to automate this labelling soon). Evaluation of the proposed work will focus on how well robots can perform suggested activities from the instructional material, repair broken processes and devices, and create processes and devices that achieve new task specifications.

4 Our behavior capture facilities

We describe our behavior capture facilities including the Motion Capture Lab (based on reflective markers and IR illumination) and the Panoptic Studio (based on several hundred video cameras), as well as virtual reality and web-based videogames, which are used to capture human teacher behavior and human-robot interaction (Figure 5).

4.1 Motion Capture Lab

The 1700 square foot Motion Capture Lab provides a resource for marker-based behavior capture of humans as well as measuring and controlling robot behavior in real time (Figure 9). It includes a Vicon Optical Motion Capture System with sixteen 200 Hz, 4Meg resolution cameras (MX-40). In addition to traditional motion capture, the Vicon system can be used in real time to track robot motion, and provide the equivalent of very high quality inertial feedback. In addition to capturing motion, we have instrumentation to capture contact forces at the hands and feet (one force gauge (IMADA DPS-44), one ATI Industrial Automation Mini85 wrist force torque sensor, and two AMTI AccuSway PLUS force plates that measure the six-axis contact force and torque at a rate of 1 kHz), and also electromyographic activity (EMG, a measure of muscle activation, Aurion ZeroWire (wireless) system

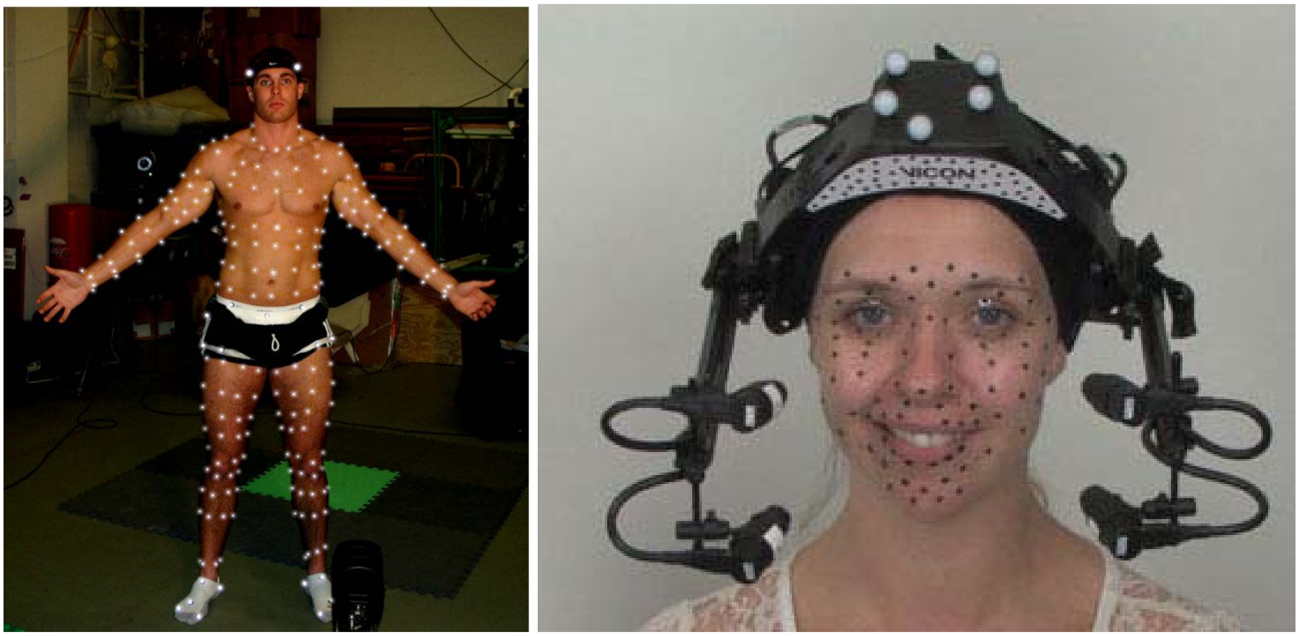


Figure 9: Capturing skin deformation as well as whole body movement

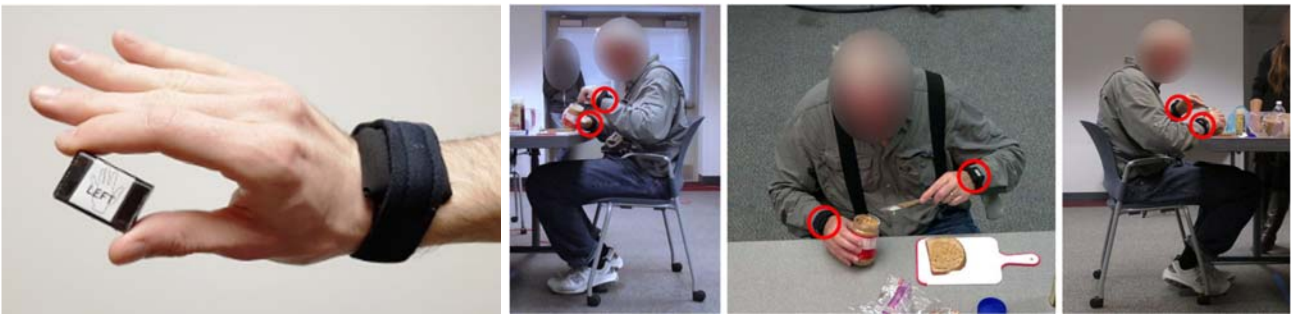


Figure 10: **Left:** A wearable accelerometer system. **Right:** A Parkinson's patient whose tremor is being monitored by cameras and wearable accelerometers (red circles).

with 16 pairs of electrodes at a rate of 5 kHz) A high-speed video camera is also used to capture skin deformation at 1 kHz. Behavior capture goes beyond motion capture with this capture of forces and muscle activation. We have also built wearable behavior capture systems (Figure 10).

4.2 Panoptic Studio

The Panoptic Studio is a multiview capture system with 521 heterogeneous sensors, consisting of 480 VGA cameras, 31 HD Cameras, and 10 Kinect v2 RGB+D sensors, distributed over the surface of geodesic sphere with a 5.49m diameter (Figure 11). The large number of lower resolution VGA cameras at unique viewpoints provide a large volume with robustness against occlusions, and allow no restriction for view direction of the subjects. The HD views provide details (zoom) of the scene. Multiple Kinects provide initial point clouds to generate dense trajectory stream.

The structure consists of pentagonal panels, hexagonal panels, and trimmed base panels. Our

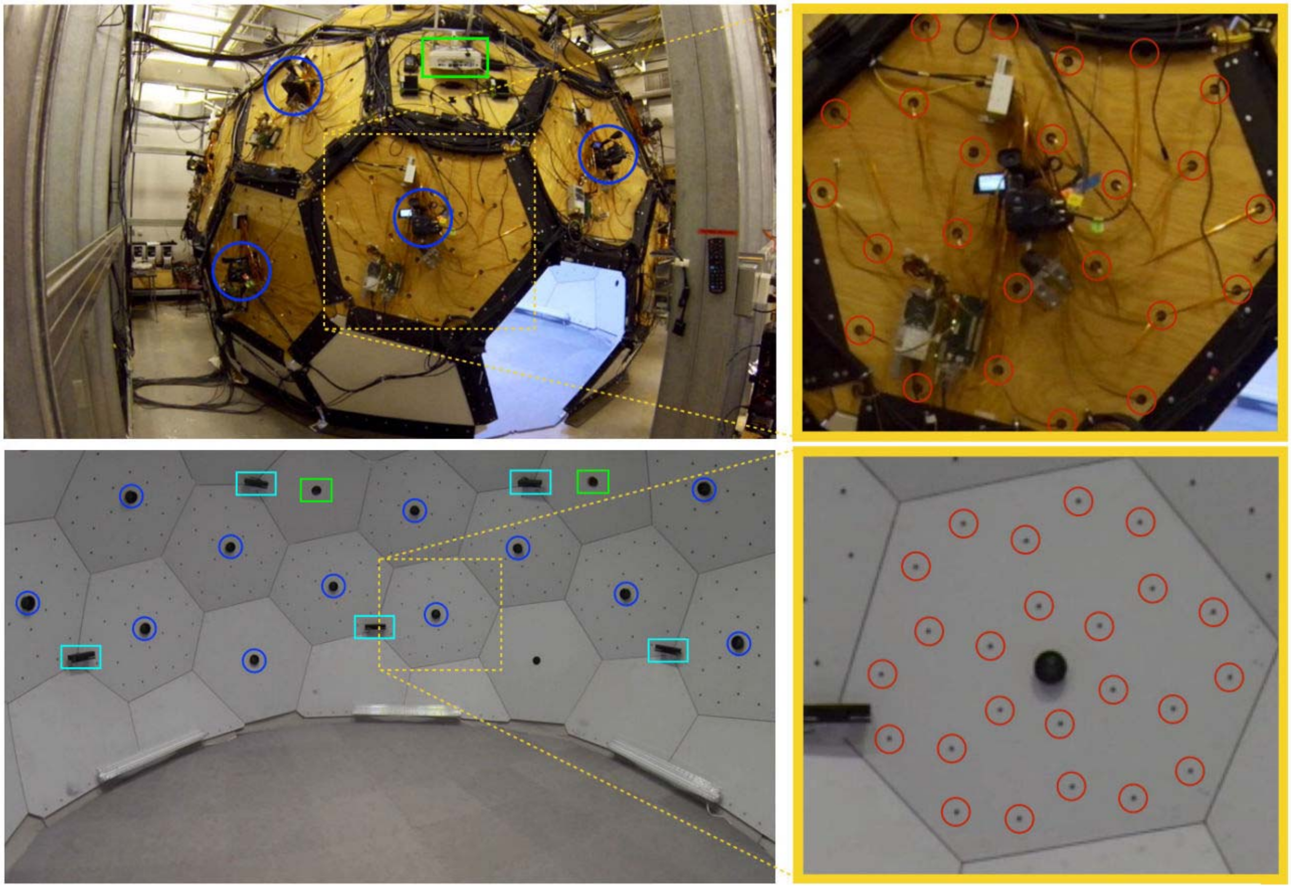


Figure 11: Panoptic Studio layout. (Top Row) The exterior of the dome with the equipment mounted on the surface. (Bottom Row) The interior of the dome. VGA cameras are shown as red circles, HD cameras as blue circles, Kinects as cyan rectangles, and projectors as green rectangles.

design was modularized so that each hexagonal panel houses a set of 24 VGA cameras. The HD cameras are installed at the center of each hexagonal panel, and projectors are installed at the center of each pentagonal panel. Additionally, a total of 10 Kinect v2 RGB+D sensors are mounted at heights of 1 and 2.6 meters, forming two rings with 5 evenly spaced sensors each.

Examples of human tracking include Figures 12, 13, and 14.

We expect to continue to make behavior capture data available on the web. We have had great success making public most data collected in our Motion Capture Lab (`mocap.cs.cmu.edu` and `kitchen.cs.cmu.edu`) and Panoptic Studio (`domedb.perception.cs.cmu.edu`). Data made available so far has been acknowledged in several hundred papers, mostly from the computer graphics, animation, and vision communities worldwide. As an example of behavior capture in the PanOptic Studio, we publicly share a novel dataset which is the largest in terms of the number of views (521 views), duration (3+ hours in total), and the number of subjects in the scenes (up to 8 subjects) for full deformable body motion capture. Our dataset is distinctive in that ours captures natural interactions of groups without controlling their behavior and appearance, and contains motions with rich social signals.

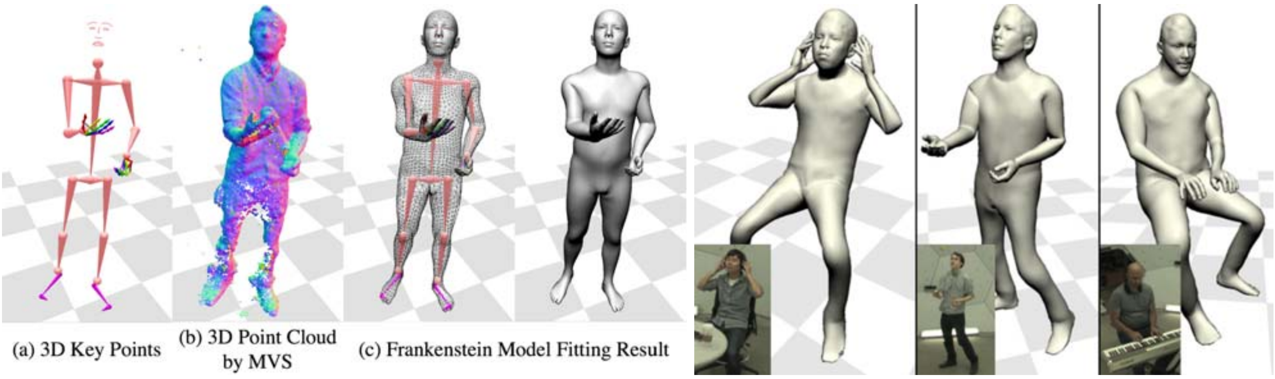

(a) 3D Key Points
(b) 3D Point Cloud by MVS
(c) Frankenstein Model Fitting Result

Figure 12: Tracking humans in the Panoptic Studio.

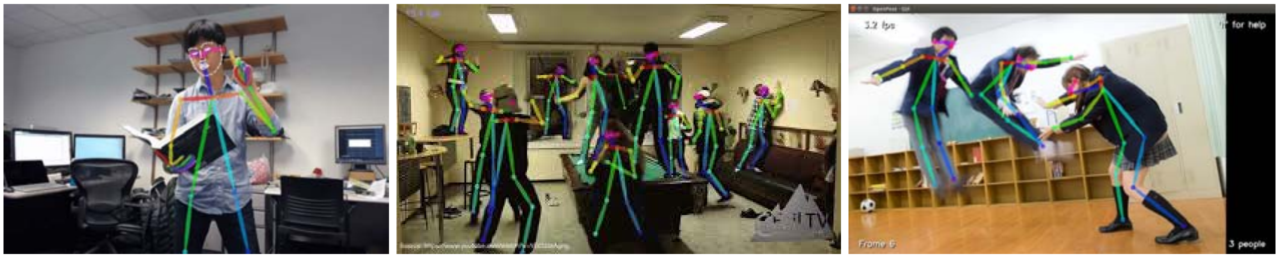


Figure 13: Openpose is widely used human tracking software that came out of the work on the Panoptic Studio.

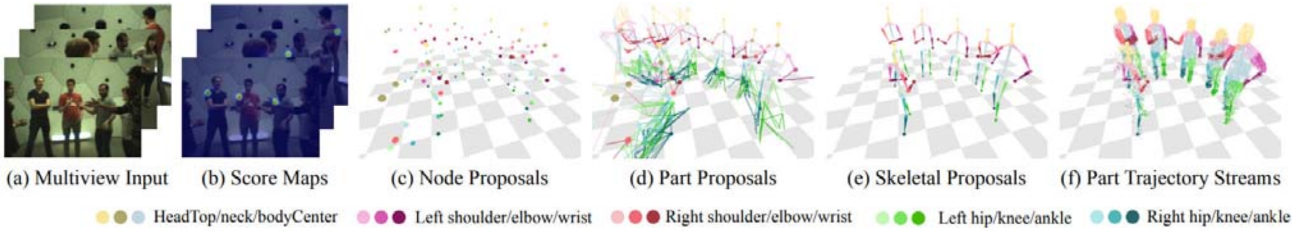

(a) Multiview Input
(b) Score Maps
(c) Node Proposals
(d) Part Proposals
(e) Skeletal Proposals
(f) Part Trajectory Streams
● ● ● HeadTop/neck/bodyCenter
● ● ● Left shoulder/elbow/wrist
● ● ● Right shoulder/elbow/wrist
● ● ● Left hip/knee/ankle
● ● ● Right hip/knee/ankle

Figure 14: Several levels of proposals generated by our method. (a) Images from up to 480 views. (b) Per-joint detection score maps. (c) Node proposals generated after non-maxima suppression. (d) Part proposals by connecting a pair of node proposals. (e) Skeletal proposals generated by piecing together part proposals. (f) Labeled 3D patch trajectory stream showing associations with each part trajectory. In (c-f), color means joint or part labels shown below the figure.

4.3 Virtual reality and videogames

We will construct virtual reality versions of “rigid body” toys and kits as well as our robots. Figure 5a shows our current virtual reality interface. Fun videogame-like simulations will be used to engage a large number of human teachers on the web for virtual teaching, and this domain is an excellent vehicle for outreach to schools and museums.

Collaboration Plan

This collaboration involves faculty from the Machine Learning Department and the Robotics Institute of the School of Computer Science (SCS) at Carnegie Mellon University. Our offices and labs are part of the same building complex. The co-PIs will collaborate on managing the project. Graduate students will be recruited from all of the departments in SCS, including the Machine Learning Department, Robotics Institute, Language Technologies Institute, Human-Computer Interaction Institute, and the Computer Science Department. This will be facilitated by the policy that any faculty member in SCS can supervise students in any other SCS department The students will be co-advised by several of the PIs encourage integration. Interaction between the PIs will be facilitated by our proximity and shared labs and students.

Roles of co-PIs: What makes this team more than the sum of the individuals is that we bring together researchers with different areas of expertise. We build on an existing machine learning and computer vision research program (Fragkiadaki), an existing humanoid robot and robot learning research program (Atkeson), a new effort in tactile sensing (Yuan), and an existing effort in machine learning and language (Mitchell). PI Katerina Fragkiadaki has worked extensively on fine-grain activity understanding and visual recognition from videos by combining semantics, geometry and unsupervised learning. Co-PI Chris Atkeson has worked extensively on robot learning, manipulation, and locomotion, as well as robot design. Co-PI Wenzhen Yuan is an expert on tactile perception, both in hardware development and algorithms for understanding tactile feedback. Co-PI Tom Mitchell has extensive experience in machine learning and natural language.

The participating investigators will work synergistically to accomplish the proposed work. Fragkiadaki, as the PI, will coordinate the efforts supported by this proposal. Her background will enable her to be a bridge between the interdisciplinary components of this proposal and support more effective interaction. She will lead work in visual perception and multimodal integration. The funding will also support several graduate students. We expect the students to allocate aspects of the project based on their interests. The PIs will supervise, guide, and provide assistance as needed, as well as focus on theoretical and algorithm development. All members of the team will work together to evaluate each other's work and results.

Project management: The co-PIs will make decisions by consensus. The PI is the final decision-maker if consensus is not reached. In addition to the co-PIs, the project includes several co-advised students. The students will interact with all of the co-PIs.

Specific collaboration mechanisms: The PIs will have multiple mechanisms in place to ensure effective collaboration. In addition to frequent ad hoc meetings and interactions, the personnel involved in this work will meet weekly to review research results, technical progress on individual technology, establishing milestones, planning for system demonstrations and software releases, and discuss future directions. We will also hold larger events such as workshops to focus attention on particular issues and get input from additional colleagues. We will share experimental setups and software repositories.

Budgetary support for collaboration: Faculty salary support and student support will provide coverage for the time necessary to interact. No additional funding is required for our regular meetings or other collaboration mechanisms. Support is requested for several students. We would ask for more but are limited by the budget limits set by the NSF. We also ask for summer support for the PIs. This support allows them to use their summer for both their individual technical contributions and the leadership of collaborative activities necessary to achieve our goals, in addition to the time provided in the academic year. Support is also requested for travel to present results at scientific conferences as well as attend PI meetings.

Timeline for the integrative activities:

Year 1 will focus on building the foundations of the project. Our goal will be to establish limited prototypes of all basic components of our learning ecosystem, including simulation and virtual reality, actual robot, and evaluation components. A major emphasis will be making basic representational decisions and developing and evaluating prototype implementations of various tasks. We will focus on integration of existing algorithms into the first version of our system. Infrastructure such as computer vision, robot control, and data collection tools will be created. We will manually build system components where learning approaches are not sufficiently mature. Another major task will be to create common software, data collection, and work sharing mechanisms that cover all of the testbeds. In terms of first year evaluation, we will implement a baseline version of the proposed approach, which will be evaluated in simulation and on our robots.

In terms of the Toys/Kits Testbed, we will continue the work we have already begun on the simplest infant toys and K-12 kits. We will build a prototype system with example perception, learned models, reasoning capabilities, behavior libraries, and learning algorithms. This example system will focus on rigid-body mechanics, so it can work with kits like the Keva Contraptions kit (balls rolling around a blocks world), as well as many kits involving marbles rolling down ramps magnetically attached to a wall, patterns of dominos falling, and other rigid-body processes. In terms of the Social/Affordance Testbed involving all robots but emphasizing the Cobots, we will integrate existing CoBot control software into our learning ecosystem. In terms of the Deformable/Liquid/Granular Testbed, we will integrate our existing food preparation software into our learning ecosystem.

In terms of general capabilities, our plan is described in the Roadmap in the description section. For example, we will begin exploring how multi-modal narrated and annotated behavior capture of humans can be used in learning from demonstration and also to define more useful component behaviors. We will develop efficient algorithms to learn and optimize temporally decomposed dynamics, including bifurcations and loops. We will develop symbolic-level reasoning to handle changes to processes, such as movement of objects, and failure of previous strategies. At the end of the first year our milestones will include simple simulations, an implementation of our prototype system on robots for rigid body mechanical processes, and evaluation infrastructure.

Year 2 will focus on exploring and implementing alternative representations, algorithms, control structures, and comparing their behavior in quantitative evaluations. We will clean up the arbitrary and expedient design choices to get something working we have already made or made in the first year. Year 2 marks the beginning of a much more thorough exploration, and a more careful statistical characterization of performance. We will seriously evaluate the Year 1 system. Space limits do not permit specific discussion of the various testbeds.

Year 3 will focus on extending our system to other domains including deformable mechanical, electrical, thermal, chemical, and combustion processes. Major emphases will be on multi-domain integration, transfer learning across domains and robots, and scaling up our knowledge bases. We will evaluate the Year 2 system, and improve the components using several forms of learning and reflection. Milestones will include the ability of the system to do 100 tasks with representative tasks in all domains. We will demonstrate learning from demonstration, simulation, practice, and reflection across a set of tasks and robots. We will seriously evaluate the Year 2 system. Space limits do not permit specific discussion of the various testbeds.

Year 4 focuses on evaluation, redesign, and refinement. Although we will iteratively formatively evaluate and refine our system throughout the duration of the project, Year 4 will focus on summative evaluation. We will also refine our algorithms in response to early evaluation results, improving the integration of components. We will evaluate our approaches both from an experimental and a theoretical point of view. Space limits do not permit specific discussion of the various testbeds.

Relation to investigators' long-term goals: The co-PIs share a goal of enabling robots to show human levels of competence in performance and learning of everyday life activities. The proposed work clearly aligns with these agendas.

Dissemination: We will setup a project web page on which we will publicly release software as well as data and experimental results of interest to the research community. The PIs will work together to organize workshops, tutorials and invited sessions at the major conferences relevant to this project. We will also include challenges of our research in the courses taught by the PIs, at the undergraduate and graduate levels. We will provide demonstrations to the large variety of visitors who come to CMU. We envision that by widely exposing our work, we will contribute to a better understanding of the functionality and benefits that our perception, learning, language, and robotics technology can bring to our society.