Information-Theoretic Reward Shaping for Multimodal Object Attribute Learning

Xiaohan Zhang
State University of New York at Binghamton
Binghamton, NY 13902, USA
Email: xzhan244@binghamton.edu

Jivko Sinapov
Tufts University
Medford, MA 02155, USA
Email: jivko.sinapov@tufts.edu

Shiqi Zhang
State University of New York at Binghamton
Binghamton, NY 13902, USA
Email: zhangs@binghamton.edu

I. Introduction

Intelligent robots are able to interact with objects through exploratory behaviors in real-world environments. For instance, a robot can take a $look$ behavior to figure out if an object is "red" using computer vision methods. However, vision is not sufficient to recognize if an opaque bottle is "full" or not, and behaviors that support other sensory modalities, such as $lift$ and $shake$, become necessary. Given the sensing capabilities of robots and the perceivable properties of objects, it is important to develop algorithms to enable robots to use multimodal exploratory behaviors to identify object properties, answering questions such as "*Is this object red and empty?*" In this paper, we use **attribute** to refer to a perceivable property (of an object) and use **behavior** to refer to an exploratory action that a robot can take to interact with the object.

Robot multimodal perception is a challenge for several reasons. First, exploratory behaviors can be costly, and even risky in the real world. For instance, to $shake$ a water bottle to identify the value of attribute "empty", the robot must first $grasp$ and $lift$ it. Those behaviors take time and can break the bottle in case of failed grasps. Second, those behaviors are not equally useful for recognizing different attributes. For instance, $lift$ is more useful than $look$ for "heavy," while $look$ works much better for "shiny." **Robot attribute learning** (RAL) algorithms aim to learn an observation model for each attribute given an exploratory behavior and play a key role in robot multimodal perception. Most existing RAL algorithms are considered **offline**: the robot learns the attributes by interacting with objects *without* considering data collection costs. In the evaluation phase, the robot uses the learned attributes to identify attributes of new objects (i.e., attribute identification). In this research, we are concerned with a novel **online** RAL (On-RAL) setting, where the robot needs to learn an action policy for interacting with objects toward efficient attribute learning and accurate attribute identification at the same time.

On-RAL faces the fundamental trade-off between exploration and exploitation. A trivial solution is to let the robot optimize its behaviors solely on attribute identification as if the attributes have been learned already. In doing so, the robot still learns the observation models of attribute-action pairs as it becomes more experienced, but this trivial solution lacks a mechanism for actively improving its long-term attribute identification performance. The main contribution of this paper is an algorithm, called *information-theoretic reward shaping* (**ITRS**), for On-RAL problems. ITRS, for the first time, equips a robot with the capability of optimizing its sequential action selection toward (efficiently and accurately) learning and identifying attributes at the same time, as shown in Figure 1. ITRS has been evaluated using two datasets: one dataset, called **CY101**, contains 101 objects with ten exploratory behaviors and seven types of sensory modalities [33]; and the other, called **ISPY32**, includes 32 objects with eight behaviors and six types of modalities [34]. Compared with existing methods from the RAL literature [2, 36], ITRS reduces the overall cost of exploration in the long term while reaching a higher accuracy of attribute identification.

The complete version of this paper has been accepted for publication under the title of "**Planning Multimodal Exploratory Actions for Online Robot Attribute Learning**" at the RSS-2021 Conference [39].

II. Related Work

Multimodal Perception in Robotics: Significant advances have been achieved recently in computer vision, e.g., [17, 26] and natural language processing, e.g., [6, 7]. While language and vision are important communication channels for robotic perception, many object properties cannot be detected using vision alone [12] and people are not always available to verbally provide guidance in exploration tasks. Therefore, researchers have jointly modeled language and visual information for multimodal text-vision tasks [25]. However, many of the most common nouns and adjectives (e.g., "soft", "empty") have a strong non-visual component [20] and thus, robots would need to perceive objects using additional sensory modalities to reason about and perceive such linguistic descriptors. To address this problem, several lines of research have shown that incorporating a variety of sensory modalities is the key to further enhance the robotic capabilities in recognizing multisensory object properties (see [4] and [19] for a review). For example, visual and physical interaction data yields more accurate haptic classification for objects [11], and non-visual sensory modalities (e.g., $audio$, $haptics$) coupled with exploratory actions (e.g., $touch$ or $grasp$) have been shown useful for recognizing objects and their properties [5, 10, 15, 22, 28], as well as grounding natural language descriptors that people

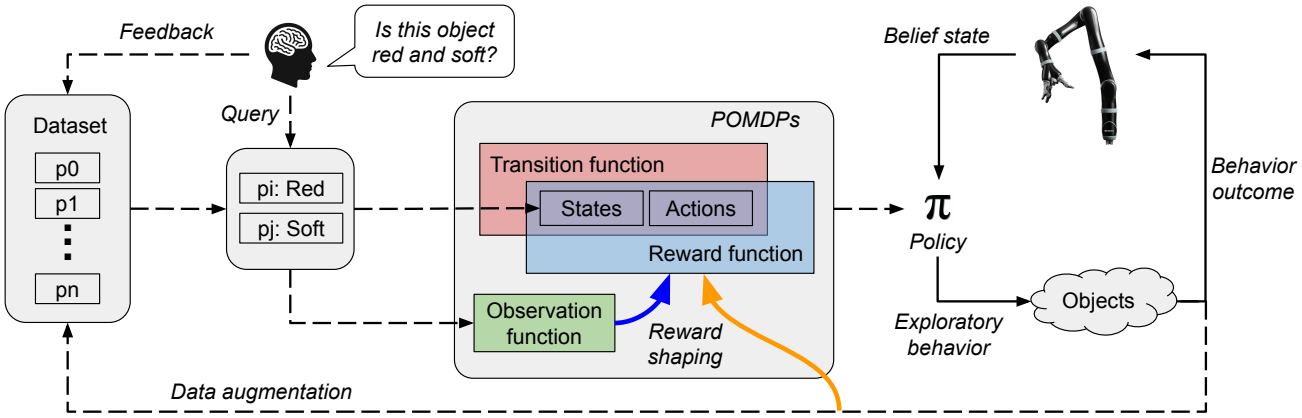


Fig. 1: An overview of the ITRS algorithm. A human user will choose an object and ask a query such as "*Is this object red and soft?*". The robot will generate a perception model on the specified attributes, i.e., "red" and "soft". Queried attributes and the corresponding perception model then will be used to construct states and the observation function of the POMDP model respectively. The reward function will be shaped by the quality of the observation function and the robot's experience. The robot uses the generated POMDP model to compute a policy π and interacts with the queried object. Newly-perceived feature data will be used to update the robot's experience and augment the dataset. Humans will give feedback to the robot's answer and attach labels to the feature data points.

use to refer to objects [3, 34]. More recently, researchers have developed end-to-end systems to enable robots to learn to perceive the environment and perform actions at the same time [18, 37].

A major limitation of these and other existing methods is that they require large amounts of object exploration data, which is much more expensive to collect as compared to vision-only datasets. A few approaches have been proposed to actively select behaviors at test time (e.g., when recognizing an object [9, 29] or when deciding whether a set of attributes hold true for an object [2]). One recent work has also shown that robots can bias which behavior to perform at training time (i.e., when learning a model grounded in multiple sensory modalities and behaviors) but they did not learn an actual policy for doing so [36]. Different from existing work, we propose a method for learning a behavior policy for object exploration that a robot can use when learning to ground the semantics of attributes.

Planning under Uncertainty: Decision-theoretic methods have been developed to help agents plan behaviors and address uncertainty in non-deterministic action outcomes [24, 32]. Existing planning models such as partially observable Markov decision process (POMDP) [13], belief space planning [23] and Bayesian approaches [27] have shown great advantages for planning robot perception behaviors, because robots need to use exploratory actions to estimate the current world state. To learn semantic attributes, robots frequently need to choose multiple actions, so POMDP which is useful for long-term planning is particularly suitable. Many of the POMDP-based robot perception methods are vision-based [8, 31, 38, 40]. Compared to those methods, our robot takes advantage of non-visual sensory modalities, such as *audio* and *haptics*.

Work closest to this research plans under uncertainty to interact with objects using multimodal exploratory actions [2], where they modeled the mixed observability [21] in domains of a robot interacting with objects (we use their work as a baseline approach in experiments). The work of Amiri et al. [2] and this work share the same spirit from the planning and perception perspectives. The main difference is that their work assumed that sufficient training data and annotations are available for the robot to learn the perception models of its exploratory actions. In comparison, we consider a more challenging setting, called "Online RAL," where the data collection and task completion processes are simultaneous.

Robot Attribute Learning (RAL): To select actions to identify objects' perceivable properties (e.g., "heavy," "red," "full," and "shiny"), robots need observation models for their exploratory actions. Researchers have developed algorithms to help robots determine the presence of possibly new attributes [16] and learn observation models of objects' perceivable properties (i.e., attributes) given different exploratory actions [34, 35, 30]. In the case where the object attributes refer to the object's function, they are then referred to as 0-order affordances [1]. Those methods focused on learning to improve the robots' perception capabilities. Once the learning process is complete, a robot can use the learned attributes to perform tasks, such as attribute identification (e.g., to tell if a bottle is "heavy" and "red"). Compared to those learning methods, we consider an online multimodal RAL setting, where the robot learns the attributes (an exploration process) and uses the learned attributes to identify object properties (an exploitation process) at the same time. The exploration-exploitation trade-off is a fundamental decision-making challenge in unknown environments. While the problem has been studied in multi-armed bandit [14] and reinforcement learning settings [32], it has not been studied in RAL contexts.

III. Algorithm Description

In On-RAL problems, the robot needs to optimize its behaviors toward not only improving the accuracy of attribute identification but also minimizing the cost of exploratory actions. We introduce two factors of *perception quality* and

interaction experience into the reward design of POMDPs to achieve the trade-off between exploration (actively collecting data for attribute learning) and exploitation (using the learned attributes for identification tasks). Intuitively, we aim to encourage the robot to select exploratory action a from those actions, where the perception model of a is of poor quality, and there is relatively limited experience of applying a to identify attribute p, i.e., the experience of (p, a) is limited.

The details of our approach are omitted from this short paper, and are available in the full paper of this work [39].

IV. Experimental Evaluation

The key hypothesis is that ITRS outperforms existing RAL algorithms in learning efficiency and task completion accuracy (there does not exist an On-RAL algorithm in the literature). Two public datasets of **CY101** [33] and **ISPY32** [34] are used in our experiments where **CY101** contains many more household objects and attributes.

A. Illustrative Trial

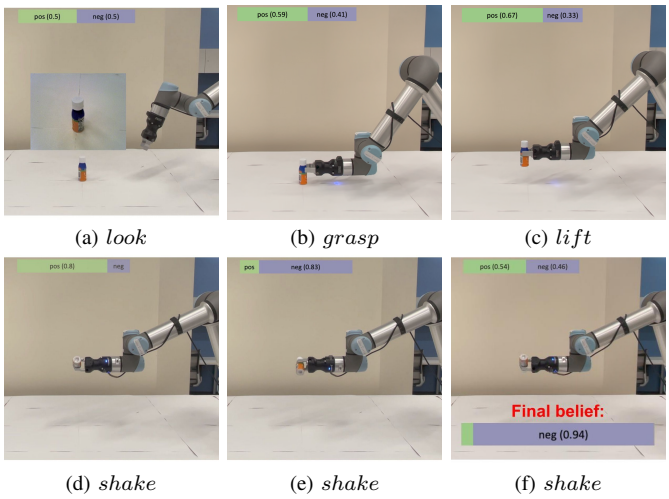
From the robot's many trials of the learning experience, we selected two trials (T_1 and T_2), where the robot faced the same object (a Coke can that has attributes "metal," "empty," and "container") and needed to answer the same question "*Is this object soft?*" From the dataset, we know that the correct answer should be "no" (the robot did not know it). T_1 appeared at the second batch of training, and T_2 appeared at the ninth. We present both trials and explain how the robot performed better in T_2.

TABLE I: Early and late observation models for action $press$

	Early phase		Late phase	
Not soft (Ground Truth)	0.68	0.32	0.82	0.17
Soft (Ground Truth)	0.50	0.50	0.20	0.80
	Not soft (Observed)	Soft (Observed)	Not soft (Observed)	Soft (Observed)

In T_1 (early learning phase), the robot first performed the $look$ action. Then, the robot had the following options: $grasp$, tap, $push$, $poke$ and $press$. Specifically, for $press$, the observation probability (shown in Table I) was nearly uniform, which is typical in the early learning phase. Among those "less useful" actions, the robot chose $grasp$. The robot's belief was changed from [0.37, 0.63] to [0.46, 0.54], where the entries represent "not soft" and "soft" respectively. After $press$, ITRS sequentially suggested $grasp$, $lift$, $hold$ and $hold$. Finally, the robot reported "positive" that resulted in a failed trial with a total cost of 55.5 seconds.

In T_2 (late learning phase), Action $press$ became more useful for identifying attribute "soft" compared to T_1, as shown in Table I. For $grasp$, the *interaction experience* was 0.67 and the observation probability was [0.66, 0,33, 0.61, 0.38] (TN, FN, FP, TP), which meant that the robot was experienced with action $grasp$ and considered $grasp$ was not as useful as $press$. Accordingly, ITRS suggested $press$ instead of $grasp$ after taking $look$. The robot's belief changed from [0.57, 0.43] to [0.67, 0.33]. After only $look$ and $press$, the robot was able to quickly report "negative", resulting in a successful trial with a total exploration cost of 22.5 seconds.



(a) $look$ (b) $grasp$ (c) $lift$

(d) $shake$ (e) $shake$ (f) $shake$

Fig. 2: A demonstration of the learned action policy using ITRS algorithm. The robot performed six actions in a row. At the beginning, the robot started with a uniform distribution (it evenly believed the object can be empty or not). After completing the six actions, the belief converged to "negative" (0.94 probability). Finally the robot selected a reporting action to report that the object is "*not empty*."

From the above two trials (same query and object in different learning phases), we see how the improved perception model of $(press, soft)$ helped the robot correctly identify "soft" with a lower cost.

B. Real Robot Demonstration

We have demonstrated the learned action policy using a real robot (UR5e arm from Universal Robots). It should be noted that the two datasets we used in this research were collected on robots that are different from the robot in the demonstration. It is a major challenge in robotics of transferring skills learned from one robot to another. To alleviate the effect caused by the heterogeneity of robot platforms, after performing each action, we sampled a data instance from **CY101**.

In the demonstration trial, our robot was given an object – a pill bottle half-full of beans. The one-attribute query was "*Is this object empty?*" Figure 2 shows a sequence of screenshots of the UR5e robot completing the task using a learned ITRS policy.

V. Conclusions

In this work, we focus on a new On-RAL problem where the robot is required to complete attribute identification tasks and, at the same time, learn its observation model for each attribute. We propose an algorithm called ITRS that selects exploratory actions toward simultaneous attribute learning and attribute identification. The proposed method and baseline methods are evaluated using two real-world datasets. Experimental results show that ITRS enables the robot to complete attribute identification tasks at a higher accuracy using the same amount of training time compared to baselines.

References

[1] Aitor Aldoma, Federico Tombari, and Markus Vincze. Supervised learning of hidden and non-hidden 0-order affordances and detection in real scenes. In *2012 IEEE International Conference on Robotics and Automation*, pages 1732–1739. IEEE, 2012.

[2] Saeid Amiri, Suhua Wei, Shiqi Zhang, Jivko Sinapov, Jesse Thomason, and Peter Stone. Multi-modal predicate identification using dynamically learned robot controllers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

[3] Jacob Arkin, Daehyung Park, Subhro Roy, Matthew R Walter, Nicholas Roy, Thomas M Howard, and Rohan Paul. Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions. *The International Journal of Robotics Research*, 39(10-11):1279–1304, 2020.

[4] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017.

[5] Raphaël Braud, Alexandros Giagkos, Patricia Shaw, Mark Lee, and Qiang Shen. Robot multi-modal object perception and recognition: synthetic maturation of sensorimotor learning in embodied systems. *IEEE Trans. on Cognitive and Developmental Systems*, 2020.

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Robert Eidenberger and Josef Scharinger. Active perception and scene modeling by planning with probabilistic 6d object poses. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1036–1043. IEEE, 2010.

[9] Jeremy A Fishel and Gerald E Loeb. Bayesian exploration for intelligent identification of textures. *Frontiers in neurorobotics*, 6:4, 2012.

[10] Dhiraj Gandhi, Abhinav Gupta, and Lerrel Pinto. Swoosh! Rattle! Thump! - Actions that Sound. In *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[11] Yang Gao, Lisa Anne Hendricks, Katherine J Kuchenbecker, and Trevor Darrell. Deep learning for tactile understanding from visual and haptic data. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 536–543. IEEE, 2016.

[12] Eleanor J Gibson. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual review of psychology*, 1988.

[13] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[14] Michael N Katehakis and Arthur F Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

[15] Matthias Kerzel, Erik Strahl, Connor Gaede, Emil Gasanov, and Stefan Wermter. Neuro-robotic haptic object classification by active exploration on a novel dataset. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[16] George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *J. of Artificial Intelligence Research*, 61:215–289, 2018.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[18] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[19] Qiang Li, Oliver Kroemer, Zhe Su, Filipe Fernandes Veiga, Mohsen Kaboli, and Helge Joachim Ritter. A review of tactile information: Perception and action through touch. *IEEE Transactions on Robotics*, 36(6): 1619–1634, 2020.

[20] Dermot Lynott and Louise Connell. Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, 41(2):558–564, 2009.

[21] Sylvie CW Ong, Shao Wei Png, David Hsu, and Wee Sun Lee. Planning under uncertainty for robotic tasks with mixed observability. *The International Journal of Robotics Research*, 29(8):1053–1068, 2010.

[22] Francisco Pastor, Jorge García-González, Juan M Gandarias, Daniel Medina, Pau Closas, Alfonso J García-Cerezo, and Jesús M Gómez-de Gabriel. Bayesian and neural inference on LSTM-based object recognition from tactile and kinesthetic information. *IEEE Robotics and Automation Letters*, 6(1):231–238, 2020.

[23] Robert Platt Jr, Russ Tedrake, Leslie Kaelbling, and Tomas Lozano-Perez. Belief space planning assuming maximum likelihood observations. 2010.

[24] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al.

Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[27] Stéphane Ross, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A bayesian approach for learning and planning in partially observable markov decision processes. *J. of Machine Learning Research*, 12(5), 2011.

[28] Amrita Sawhney, Steven Lee, Kevin Zhang, Manuela Veloso, and Oliver Kroemer. Playing with food: Learning food item representations through interactive exploration. In *Proceedings of 17th International Symposium on Experimental Robotics (ISER '20)*. Springer Proceedings in Advanced Robotics (SPAR), November 2021.

[29] Jivko Sinapov, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 62(5): 632–645, 2014.

[30] Jivko Sinapov, Priyanka Khante, Maxwell Svetlik, and Peter Stone. Learning to order objects using haptic and proprioceptive exploratory behaviors. In *IJCAI*, pages 3462–3468, 2016.

[31] Mohan Sridharan, Jeremy Wyatt, and Richard Dearden. Planning to see: A hierarchical approach to planning visual actions on a robot using POMDPs. *Artificial Intelligence*, 174(11):704–725, 2010.

[32] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[33] Gyan Tatiya and Jivko Sinapov. Deep multi-sensory object category recognition using interactive behavioral exploration. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7872–7878. IEEE, 2019.

[34] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multi-modal grounded linguistic semantics by playing "I Spy". In *IJCAI*, pages 3477–3483, 2016.

[35] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J Mooney. Opportunistic active learning for grounding natural language descriptions. In *Conference on Robot Learning*, pages 67–76. PMLR, 2017.

[36] Jesse Thomason, Jivko Sinapov, Raymond Mooney, and Peter Stone. Guiding exploratory behaviors for multimodal grounding of linguistic descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[37] Chen Wang, Shaoxiong Wang, Branden Romero, Filipe Veiga, and Edward Adelson. SwingBot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pages 5633–5640, 2020.

[38] Shiqi Zhang, Mohan Sridharan, and Christian Washington. Active visual planning for mobile robot teams using hierarchical POMDPs. *IEEE Transactions on Robotics*, 29(4):975–985, 2013.

[39] Xiaohan Zhang, Jivko Sinapov, and Shiqi Zhang. Planning multimodal exploratory actions for online robot attribute learning. In *Proceedings of the Robotics: Science and Systems (RSS) Conference*, 2021.

[40] Kaiyu Zheng, Yoonchang Sung, George Konidaris, and Stefanie Tellex. Multi-resolution POMDP planning for multi-object search in 3d. *arXiv preprint arXiv:2005.02878*, 2020.