Luke E. Richards and Cynthia Matuszek. "Learning to Understand Non-Categorical Physical Language for Human Robot Interactions." In *Proceedings of the RSS 2019 workshop on AI and Its Alternatives in Assistive and Collaborative Robotics (RSS: AI+ACR).*

Learning to Understand Non-Categorical Physical Language for Human Robot Interactions

Luke E. Richards and Cynthia Matuszek
University of Maryland, Baltimore County – Baltimore, Maryland 21042
lurich1,cmat@umbc.edu

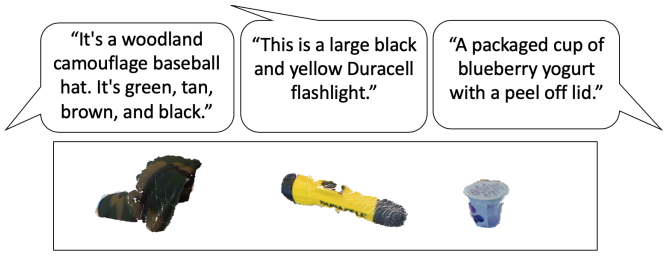
***Abstract*—Learning the meaning of language with respect to the physical world in which a robot operates is a necessary step for shared autonomy systems in which natural language is part of a user-specific, customizable interface. We propose a learning system in which language is grounded in visual percepts without pre-defined category constraints by combining CNN-based visual identification with natural language labels, moving towards making it possible for people to use language as a high-level control system for low-level world interactions, allowing a system to operate on shared visual/linguistic embeddings. We evaluate the efficacy of this learning by evaluating against a well-known object dataset, and report preliminary results that outline the feasibility of pursuing a visual feature approach to domain-free language understanding.**

I. Introduction

One of the core questions of shared-autonomy systems is the nature of interfaces and communications between human and robotic partners. Possible interfaces range from those which are purely responsive to user intent, through traditional tablet- or screen-based interfaces, to interfaces which use natural language dialog or spoken interactions [26]. Humans have spent years defining and redefining language, our major interface with one another. This positions natural language as an intuitive interface for collaborating with robots. However, although language is well-positioned to support shared autonomy in theory, in practice, understanding natural language in the general sense is very much an open problem.

Nonetheless, many of the tasks that may be of interest to such a system involve simpler physical tasks, such as identifying and interacting with objects in the environment. An additional complexity is that language use is frequently idiosyncratic, requiring efficient user-specific learning. For this, a major relevant area of research is *grounded language* [18], in which models of the environment, language semantics, and sometimes user intent are jointly learned in order to understand language in the context of a robot's physical, sensed environment. People's language use and preferred level of control is idiosyncratic, and learning a customized user model is an important element of user satisfaction [9].

Much of the existing work on grounded language understanding learns in well-defined categories, e.g., learning shapes or colors, or learning to understand action commands in a constrained context. We explore how language can be used in interfacing with a robotic agent in learning how users may naturally refer to possibly unfamiliar objects. We use CNN-based visual feature understanding paired in a joint-probability grounded language model to ground language from user given descriptions to household objects (see fig. 1) without defining categories of language that can be learned. Our primary contribution is relaxing these constraints by using multimodal deep learning to understand how language refers to objects in an environment regardless of their category. Preliminary results over a popular and openly available RGB-D dataset suggest the effectiveness of this approach.



cap= 1.0	black= 1.0	container= 1.0
green= 0.28	duracell= 0.99	yogurt= 0.99
hat= 0.99	electric= 1.0	white= 0.44
...	...	...

Fig. 1: Examples of object descriptions from Amazon Mechanical Turk workers above RGB point clouds. The table below gives classifier outputs for selected words.

II. Related Work

A. Shared Autonomy Interfaces

Shared autonomy between a human and a robot often requires some amount of explicit or implicit communication, meaning an interface of some kind is required. Possible approaches range from having responsive "mechanically transparent" [1] interactions, for example using teleimpedance to take direction from a user [8]; to interfaces where muscle [10] or brain [13] signals are monitored and used to adjust a robot's actions, but users have access to feedback showing their muscular input; to systems that incorporate direct user input, for example using a tablet [2] or an interactive dialog [32].

Natural language is particularly well-suited for supporting shifting levels of abstraction, as may be appropriate for adjustable autonomy and mixed-initiative control systems [23]. This is particularly true in physical settings where a person's understanding of context may improve shared perfor-

mance [32]. Our work is most similar to that of Scalise et al. [26], in which non-constrained language is used to interact with the robot, but differs in that their goal is to give manipulation instructions in a simulated environment, whereas the work described here is focused on learning language about real objects from complex, real-world sensor data.

B. RGB-D Object Recognition and Analysis

Object recognition is a wide-ranging area of research. In our work we focus on a multimodal combination of RGB and depth images; in this section we briefly describe the most related work in this space using RGB-D data. Early methods of RGB-D object recognition focused on extracting features for separate categories such as gradient, color, and shape [3]. While current methods deploy advances in deep learning.

Extracting features with hierarchical matching pursuit (HMP) introduces an unsupervised feature extraction network, which allows a model to learn high level features through a layered approach to combining RGB and depth images. This method was early work in the concept of gaining visual features that would be descriptive of the entire object itself.

Transfer learning combined with deep learning has been a major catalyst in the success of computer vision tasks by introducing *transferable layers* in vision models. These initial layers can then specialize to a task in a specific domain. This concept has been particularly popular due to the large-scale dataset ImageNet [6] being used to learn generalized concepts between vision tasks. While the ImageNet dataset [6] has furthered work in RGB object recognition, the benefits are yet to be fully explored in the RGB-D space. Particularly, the concept of a single GPU neural network that can be shared and trained, CaffeeNet [11], led to even more success in the space. Eitel et al. [7] introduced a method that combines the benefits of transfer learned RGB models to both RGB and depth images for object recognition. In our work, we use this state-of-the-art model to extract visual features (see III-B for details).

C. Language in HRI

Language is used to communicate, refer to, and describe the physical world. The intuitive idea to allow robotic agents to comprehend and use natural language in their operations is encapsulated in a multitude of work. Grounded language learning is the concept of learning the groundings of language to perception [5]. Grounding language has been an active field in the intersection of language and vision communities. Projects such as image caption generation and recognition [14, 24] and text-to-image synthesis [33] showcase the joint interest between communities.

When this interest moves into the physical world using robotic agents, the perceptual input that language can be grounded to increases. Language can be grounded to manipulation tasks [25, 27], navigation tasks [20, 30], and assistive robotics [4]. In all theses tasks, there is a need to understand the referent language (nouns, adjectives, and more) that is aligned with objects that occupy the physical spaces.

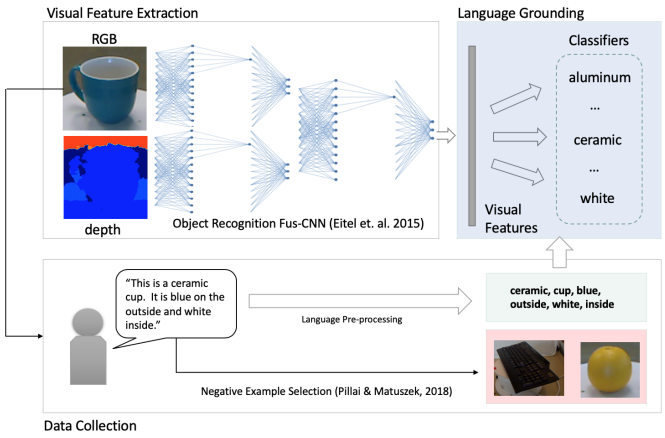


Fig. 2: Our proposed domain-free model using the visual features from object recognition system [7], creating word-as-classifier models. This method fuses two CNN architectures for RGB and depth images into fully connected fusion layers. We remove the softmax layer from their approach, exposing rich multimodal features for learning groundings.

Work in learning models for color, shape, object, haptics, and sound with predefined unique feature channels [22, 19, 31] have resulted in successful groundings. However, our work explores using a set of general features to learn groundings outside of predefined feature channels. While datasets of images and rich natural language aligned with those images are increasing, RGB images with depth sensor (RGB-D) data create a new learning paradigm for grounded language learning, as well as a need for rich datasets to benchmark future works on. Sun et al. [29] introduced the first steps towards our dataset by aligning the sensor data with user given attributes. Our work uses the same sensor dataset, however, we use full-sentence natural language rather than single word attributes in categories.

III. Approach

In this section, we introduce the process of obtaining natural language descriptions paired with RGB-D sensor data through Amazon Mechanical Turk. We then detail how we modify the robust RGB-D object identification approach of Eitel et al. [7] to extract visual features. Finally, we outline how we use a joint learning objective that combines both visual and language features to train word-as-classifier models (see fig. 2 for an overview). This approach allows the system to learn how a person might refer to objects in the environment at a semantic level, leaving low-level classification and action details to a hypothetical robot assistant.

A. Data Collection

We use the well-known UW RGB-D object set [16, 17], which includes roughly 40,000 RGB-D images of 300 objects in 51 categories. This dataset includes point clouds as well as images and masks. We select five images of each object

using stratified random sampling, giving us a sample of 1,500 RGB-D images with an unfixed collection of angles from each object. Images were then uploaded to Amazon Mechanical Turk, where workers gave short descriptions of each object as if they were speaking to another person. In order to obtain diverse descriptions, we avoided using language that would 'prime' workers to describe objects in a particular way. Workers were also encouraged not to describe the picture itself (such as "the photo is blurry" or "the photo has a red cap").

This resulted in a total of 8,186 unprocessed description responses. While some workers provided incomplete sentences or described the photo rather than the object, the majority offered rich contextual language about the images. Only descriptions which clearly did not follow instructions were removed, reducing the total number of descriptions to 7,455 complete sentence descriptions of 300 objects, giving almost 25 per object. (This number is smaller than expected because there were some images workers were unable to give a description for, either because they were too blurry, or too hard to parse visually out of context.)

B. Visual Features

Our approach to extracting visual features is drawn from the robust object recognition method of Eitel et al. [7]; however, in our results, we demonstrate that with minor modifications, this approach can be used to extract features suitable for understanding a user's high-level language in a variety of categories (see fig. 1).

Broadly, artificial neural networks (ANNs) allow for high dimensional inputs to be condensed to meaningful representations of features in the data. Due to the nature of neural networks, the final layer offers high-level features for the objects. This is true of Convolutional Neural Networks (CNNs), especially for extracting useful features in object recognition tasks [15]. While many, such as the method of Eitel et al. [7], employ a softmax function to perform classification, removing the softmax function exposes rich features that can be used for our grounded language task.

The network consists of two seven layer CNNs, one per sensor type, that combine into two final fusion layers. The final layer of the network allows for 51 features to be extracted from the joint networks of the RGB and depth images. We extract these features for each of the 1,500 images sampled from the RGB-D object dataset paired with the natural language descriptions given from workers on Amazon Mechanical Turk. Once the visual features and language is paired, we start the grounded language model learning process.

C. Category-Free Joint Language Learning

We extend the grounded language learning system of Pillai and Matuszek [22]. The basis of this work is a joint model combining perception and language models [19] to learn language groundings. For this model, the groundings are learned solely from dataset itself. No prior representations of the objects nor the language is required. In this system, a word-as-classifier approach is taken, meaning each token has a single binary classifier trained to predict whether an object is described by that word. In this infrastructure, *"red"-as-classifier* would classify a red apple and a green apple as positive and negative, while *"apple"-as-classifier* would classify both as positive.

While previous work was constrained by the domains of language that could be learned, such as shape, color, and object, we seek to use multipurpose visual features from the method described in III-B to introduce end-user abstraction. Our system therefore learns from a single source of features rather than separate domain-specific sets. Another change is learning by image rather than by object instance. This preserves visual differences in the object's orientation and appearance, so the system has the opportunity to learn language relevant to only some images. (For example, from some angles the baseball cap in fig. 1 might look like a hemisphere, whereas in the image it does not.)

We aggregate all descriptions given by workers to each image, creating a more exhaustive 'descriptive document' for each image. These descriptive documents are preprocessed to remove singletons and stop words. Visual features (extracted per image) are paired with documents aligned with that image. We consider an instance as a positive training example of the token if that token occurs more than once in the description document, and choose negative examples by document semantic vector distance. We then train a logistic regression model for each token.

IV. Preliminary Results

We train our word-as-classifier models by splitting our dataset described in III-A into a training set and testing set using four-fold cross validation. We report the preliminary results of these tests averaged over 25 runs with random splits in order to avoid outlier results. The highest-level average F1 score is 0.689, with recall of 0.609 and precision of 0.903. These values are consistent with previous work on grounding language, but do not depend on using category-specific features, and in fact learn the meanings of words in several categories with high precision (see fig. 3).

The RGB-D dataset follows WordNet [21] in creating hierarchical structures, such as a potato being a "root vegetable" and "vegetable." While these categories are defined in WordNet, there are further subcategories that can be learned in our dataset; for example, users defining some "food cups" as "yogurt." While we do not create hierarchical representations, unlike Sun and Fox [28], our approach does allow for subcategories to be defined with the abstraction of user-defined language groundings. Other examples include "football" and "soccer," which classify different objects in the dataset.

To demonstrate the results of our approach, we show some examples from various domains in fig. 3. The 'domain' (category) of each classifier was assigned by examination, and words were selected to provide representative breadth of coverage, for both well-performing and weakly performing classifiers. We examine tokens in our model that exemplify the potential for the category-free model. Classifiers describing

token	domain	precision	recall	f1
ball	Shape	1.00	0.62	0.73
black	Color	0.79	1.00	0.86
blue	Color	0.96	0.57	0.68
bound	Adjective	1.00	0.60	0.71
box	Shape	1.00	1.00	1.00
brown	Color	1.00	0.89	0.93
ceramic	Material	1.00	0.61	0.72
chips	Subcategory	1.00	0.60	0.72
cleaning	Action	1.00	0.61	0.72
digital	Adjective	1.00	0.62	0.73
green	Color	1.00	1.00	1.00
hair	Noun	1.00	0.62	0.73
handle	Component	1.00	0.62	0.72
juice	Noun	0.96	0.65	0.74
light	Shade	1.00	0.89	0.94
orange	Color/Object	1.00	1.00	1.00
pink	Color	1.00	0.61	0.72
red	Color	0.81	1.00	0.88
top	Location/Noun	1.00	0.60	0.72
tube	Shape	1.00	0.59	0.71
white	Color	0.97	1.00	0.98
yellow	Color	0.89	1.00	0.93

Fig. 3: Example tokens from the test set with their given performance metrics. Domains are assigned by examination.

size, such as "small," were learned unexpectedly well, a product of the fixed camera distance from objects in the dataset. Classifiers for shape words, such as "box," 'ball," and "tube," performed well, although this may be partly due to overfitting to the specific object types in the dataset (tubes of toothpaste).

Tokens associated with colors and non-object specific language were able to be learned, and performance was consistent with category constrained methods [29, 12]. This finding sheds light on whether these features are broad enough to be used in the context of grounded language learning. While the dataset provided little shape specific language, we see initial findings to suggest this to be a viable method for extracting rich visual features to support grounded language learning.

V. Conclusion

We present a grounded language learning system suitable for supporting user-specific, language-based human-robot interfaces. We employ the well-known object recognition system of Eitel et al. [7] to extract rich visual features for an intuitive, category-free joint model grounded language learning system, and introduce a dataset of natural language aligned with a popular real-world sensor dataset. A series of classifiers denoted by descriptions are trained and evaluated on a held-out data set. Our initial results support the theory that category-free language learning is both feasible and desirable. In future work, we intend to explore more sophisticated language models, such as, semantic parsing to further the information provided from the natural language descriptions. We plan to use the insights from this work in exploring multimodal object embeddings in pursuit of furthering the work in category-free grounded language learning.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1637614 and No. 1657469.

References

[1] Victoria Alonso and Paloma de la Puente. System transparency in shared autonomy: A mini review. *Frontiers in neurorobotics*, 12, 2018.

[2] Peter Birkenkampf, Daniel Leidner, and Christoph Borst. A knowledge-driven shared autonomy human-robot interface for tablet computers. In *IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014.

[3] Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. Object recognition with hierarchical kernel descriptors. In *Computer Vision and Pattern Recognition*, 2011.

[4] Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder, and Brian Scassellati. Situated human-robot collaboration: predicting intent from grounded natural language. *International Conference on Intelligent Robots and Systems (IROS)*, pages 827–833, 2018.

[5] David L. Chen and Raymond J. Mooney. Learning to sportscast: a test of grounded language acquisition. In *International Conference on Machine Learning (ICML)*, 2008.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[7] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin A. Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust RGB-D object recognition. *International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, 2015.

[8] Simone Fani, Simone Ciotti, Manuel G Catalano, Giorgio Grioli, Alessandro Tognetti, Gaetano Valenza, Arash Ajoudani, and Matteo Bianchi. Simplifying telerobotics: wearability and teleimpedance improves human-robot interactions in teleoperation. *IEEE Robotics & Automation Magazine*, 25(1):77–88, 2018.

[9] Deepak Gopinath, Siddarth Jain, and Brenna D Argall. Human-in-the-loop optimization of shared autonomy in assistive robotics. *IEEE Robotics and Automation Letters*, 2(1):247–254, 2016.

[10] Siddarth Jain, Ali Farshchiansadegh, Alexander Broad, Farnaz Abdollahi, Ferdinando Mussa-Ivaldi, and Brenna Argall. Assistive robotic manipulation through shared autonomy and a body-machine interface. In *2015*

IEEE international conference on rehabilitation robotics (ICORR), pages 526–531. IEEE, 2015.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[12] Caroline Kery, Francis Ferraro, and Cynthia Matuszek. ¿Es un plátano? exploring the application of a physically grounded language acquisition system to Spanish. In *Proc. of the Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 7–17, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Hyun K Kim, J Biggs, W Schloerb, M Carmena, Mikhail A Lebedev, Miguel AL Nicolelis, and Mandayam A Srinivasan. Continuous shared control for stabilizing reaching and grasping with brain-machine interfaces. *IEEE Transactions on Biomedical Engineering*, 53(6):1164–1173, 2006.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60:84–90, 2012.

[16] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *2011 IEEE international conference on robotics and automation*, pages 1817–1824. IEEE, 2011.

[17] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. RGB-D object recognition: Features, algorithms, and a large scale benchmark. In *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*, pages 167–192, 2013.

[18] Cynthia Matuszek. Grounded language learning: Where robotics and nlp meet. In *IJCAI*, pages 5687–5691, 2018.

[19] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012.

[20] Cynthia Matuszek, Evan Herbst, Luke S. Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *ISER*, 2012.

[21] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[22] Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA, 2018.

[23] Benjamin Pitzer, Michael Styer, Christian Bersch, Charles DuHadway, and Jan Becker. Towards perceptual shared autonomy for robotic mobile manipulation. In *2011 IEEE International Conference on Robotics and Automation*, pages 6245–6251. IEEE, 2011.

[24] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.

[25] Achyutha Bharath Rao, Krishna Krishnan, and Hongsheng He. Learning robotic grasping strategy based on natural-language object descriptions. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 882–887. IEEE, 2018.

[26] Rosario Scalise, Yonatan Bisk, Maxwell Forbes, Daqing Yi, Yejin Choi, and Siddhartha Srinivasa. Balancing shared autonomy with human-robot communication. *arXiv preprint arXiv:1805.07719*, 2018.

[27] Lanbo She and Joyce Yue Chai. Interactive learning of grounded verb semantics towards human-robot communication. In *ACL*, 2017.

[28] Yuyin Sun and Dieter Fox. NEOL: Toward never-ending object learning for robots. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1621–1627, 2016.

[29] Yuyin Sun, Liefeng Bo, and Dieter Fox. Attribute based object identification. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2096–2103. IEEE, 2013.

[30] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[31] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. Improving grounded natural language understanding through human-robot dialog, 2019.

[32] Thomas Witzig, J Marius Zöllner, Dejan Pangercic, Sarah Osentoski, Rainer Jäkel, and Rüdiger Dillmann. Context aware shared autonomy for robotic manipulation tasks. In *International Conference on Intelligent Robots and Systems*, pages 5686–5693. IEEE, 2013.

[33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan R. Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.