A Testbed for Learning by Demonstration from Natural Language and RGB-Depth Video

Young Chol Song and **Henry Kautz**
Department of Computer Science
University of Rochester
Rochester, NY 14627
(585) 275-5671
{ysong,kautz}@cs.rochester.edu

Abstract

We are developing a testbed for learning by demonstration combining spoken language and sensor data in a natural real-world environment. Microsoft Kinect RGB-Depth cameras allow us to infer high-level visual features, such as the relative position of objects in space, with greater precision and less training than required by traditional systems. Speech is recognized and parsed using a "deep" parsing system, so that language features are available at the word, syntactic, and semantic levels. We collected an initial data set of 10 episodes of 7 individuals demonstrating how to "make tea", and created a "gold standard" hand annotation of the actions performed in each. Finally, we are constructing "baseline" HMM-based activity recognition models using the visual and language features, in order to be ready to evaluate the performance of our future work on deeper and more structured models.

Most research in AI has explored problems of natural language understanding, visual perception, and learning and reasoning with commonsense knowledge in isolation. Recently, however, a number of researchers have argued that such a "divide and conquer" approach has reached a point of diminishing returns, and that significant progress in any of the areas requires a more integrated approach. In work coming out of the natural language community, this new direction[1] has been called grounded language learning (Branavan, Zettlemoyer, and Barzilay 2010; Kollar et al. 2010; Chen and Mooney 2011), while in the machine vision community people speak of high-level or knowledge-based scene understanding (Kembhavi, Yeh, and Davis 2010). It is quite challenging, however, to begin this kind of research on integrated intelligence for two significant reasons: first, the work would appear to require expertise in (at least) natural language processing, vision, and knowledge representation and reasoning; and second, it is difficult to quantitatively measure progress, because there are few if any data sets and previous approaches against which one can compare.

Our project addresses both of these concerns, and also forms a foundation for our own future work on integrated intelligent agents that can be taught to recognize and assist with complex real-world tasks. We are developing a testbed for learning by demonstration from natural language and sensor data, with the initial domain of kitchen activities (Swift et al. 2012).

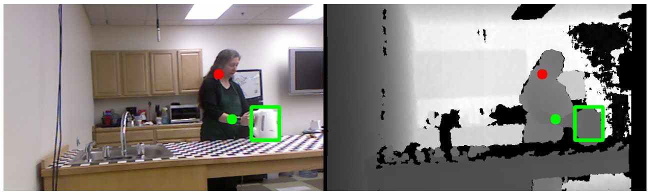

Figure 1: A screenshot from the activity recognition dataset, showing the RGB and depth streams. Agent, hand and object locations are marked.

This part of our project is similar to what is being done in the CMU's Quality of Life Grand Challenge Data Collection (la Torre et al. 2009), which also records and annotates video and audio of kitchen activities. However, in addition to the raw data and highest-level annotations, we are extracting generally useful and meaningful semantic visual features from the data streams (*e.g.* "left hand is directly above cup"), and both syntactic and semantic language features, including a parse tree and logical form for each utterance. We believe that the inclusion of features at this level of abstraction will broadly support work by researchers in machine learning and knowledge representation and reasoning. A second important difference from the CMU data set is that we are capturing 3D point-cloud data. Many problems of object segmentation and localization that are beyond the state of the art can be solved in a straightforward manner with RGB-D data, while still providing the raw point-cloud data for machine vision oriented researchers. Finally, our collection differs from that of the CMU set in that the audio is of the subject deliberately describing the steps of the activity being performed. This makes the language data appropriate for research on learning by demonstration.

Feature Extraction

Visual Features

Our visual recognition system recognizes the following high-level features: 1) object and agent location, 2) hand

[1] Or more accurately, a very old direction (Winograd 1972)

placement and interaction with objects, and 3) spatial relation of objects with respect to each other (*e.g.* above, directly-above, co-planar, etc). Point-cloud extraction using the depth stream combined with color histogram classification from the RGB stream is used to determine the location of the objects and agent. Skin detection using a Gaussian mixture model is used to detect hands and then is associated with objects in the scene.

Language Features

The language features are extracted from parses of the transcriptions of the natural language input as follows. The audio from each tea-making session was transcribed by hand for the gold standard. The transcriptions were then parsed with the TRIPS parser (Allen, Swift, and de Beaumont 2008), which uses a semantic lexicon and ontology to create a logical form (LF) that includes thematic roles, semantic types, and semantic features, yielding richer representations than "sequence of words" models. The LFs were then processed with the TRIPS Interpretation Manager (IM) to extract a concise event description from each clause derived from the main verb and its arguments. The event descriptions are formulated in terms of the semantic types in the LF and consist of short phrases such as **CREATE TEA**, **CLOSE LID**, and **POUR WATER INTO CUP**. These were used as language features in the model. The IM also performs reference resolution on referring expressions, and we plan to incorporate this information in future work.

Baseline Models

We are also creating several "baseline" models for the specific task of learning and recognizing sequences of activities. These baseline models do not begin to exhaust the range of learning, representation, and reasoning tasks that our domain and data can support, such as learning object categories or learning the full range of structures (*e.g.*, hierarchy, sensing, conditional execution, repetition) that appear in natural activities. However, baseline models are vital if one wishes to determine if more complex models actually solve a harder problem, or only solve an easy problem in a difficult way. Our baseline models are varieties of Hidden Markov Models (HMM). These include a simple HMM with one state per activity, and a hierarchical HMM, where a bank of independent HMMs, one per activity, provide virtual evidence to a higher-level HMM. Here, we describe some results on the performance of our first baseline model, where we consider the activities in the sequence to be values of a hidden state and vision and language features previously described as direct observations.

We conducted an evaluation using five sequences from our annotated activity dataset, where two sequences were from the same participant. Table 1 shows the percentage of correctly identified activities using our baseline method, where we divided results from within and between subjects. Despite using a small number of training examples and a simplified model, we see that within subject recognition when using both vision and language features show promising results. We expect using more advanced features and models will improve our recognition rates even further.

	Vision	Language	Both
Within subject	45%	19%	54%
Between subject	18%	34%	8%

Table 1: Percentage of correctly identified sequence with the baseline model using only vision/language features or both

Future Work

We provide a testbed for recognizing activities using vision and language in an instructive setting. From an activity recognition perspective, we hope to extend our baseline to incorporate more complicated models such as using Markov Logic in detecting activities over time (Brendel, Fern, and Todorovic 2011). Furthermore, we plain on investigating deeper models of activity learning and recognition; new models that draw on both classic work finding "minimal explanations" in plan recognition (Allen et al. 1991) and more recent work on probabilistic graphical models for symbol grounding (Tellex et al. 2011).

Acknowledgements

This work was supported in part by NSF grants IIS-1012205 and IIS-1012017, "Activity Recognition and Learning for a Cognitive Assistant."

References

Allen, J.; Kautz, H.; Pelavin, R.; and Tennenberg, J. 1991. *A Formal Theory of Plan Recognition and its Implementation*. Morgan Kaufmann.

Allen, J. F.; Swift, M.; and de Beaumont, W. 2008. Deep semantic analysis of text. In *Proc. of the Conf. on Semantics in Text Processing*.

Branavan, S. R. K.; Zettlemoyer, L. S.; and Barzilay, R. 2010. Reading between the lines: Learning to map high-level instructions to commands. In *Proc. of ACL 2010*.

Brendel, W.; Fern, A.; and Todorovic, S. 2011. Probabilistic event logic for interval-based event recognition. In *Proc. of CVPR 2011*.

Chen, D. L., and Mooney, R. J. 2011. Learning to interpret natural language navigation instructions from observations. 859–865.

Kembhavi, A.; Yeh, T.; and Davis, L. 2010. Why did the person cross the road (there)? scene understanding using probabilistic logic models and common sense reasoning. In *Proc. ECCV*.

Kollar, T.; Tellex, S.; Roy, D.; and Roy, N. 2010. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*.

la Torre, F. D.; Hodgins, J.; Montano, J.; Valcarcel, S.; Forcada, R.; and Macey, J. 2009. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Technical Report CMU-RI-TR-08-22.

Swift, M.; Ferguson, G.; Galescu, L.; Chu, Y.; Harman, C.; Jung, H.; Perera, I.; Song, Y. C.; Allen, J.; and Kautz, H. 2012. A multimodal corpus for integrated language and action. In *Proc. of the Int. Workshop on MultiModal Corpora for Machine Learning*.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S. J.; and Roy, N. 2011. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine* 32(4).

Winograd, T. 1972. *Understanding Natural Language*. Academic Press.