KILLE: Learning Objects and Spatial Relations with Kinect

Erik de Graaf
University of Gothenburg
gusdegrer@student.gu.se

Simon Dobnik
University of Gothenburg
simon.dobnik@gu.se

Abstract

We present a situated dialogue system designed to learn objects and spatial relations from relatively few examples, based on camera imagery and dialogue interaction with a human partner. We also report on the baseline evaluation of the system.

1 Introduction

Grounding, the linking of real world objects and situations involving objects to their computational semantic representations, is a necessary step for meaningful interaction with robots (Roy, 2005). Systems that operate within the real world will often encounter novel situations and word usages and therefore they will need to *learn* new semantic representations. In contrast to state of the art systems that work with large databases of images to learn from, our system tries to learn grounded meanings of objects and spatial relations from a very few examples presented to the system in situated interactive learning. Our long term goal is to investigate how various dialogue interaction strategies with a human can leverage the sparsity of observable data.

2 Object and scene recognition

The hardware used is a Kinect 3d camera, connected to a computer. The camera is mounted stationary to a table on and over which objects are presented to the system. The Freenect drivers[1] are used to capture data from the camera and to forward them to the Robot Operating System (ROS) framework (Quigley et al., 2009). The dialogue is managed by OpenDial (Lison, 2014), including speech recognition and speech synthesis. Rules for the dialogue system are written in OpenDial's own XML format. Objects are learned by storing the recognized SIFT features or SIFT descriptors (Lowe, 2004) of each object instance that are calculated from the frames the camera forwards. Before learning and recognizing objects the background is removed. This way we remove distracting features not belonging to the object in focus. SIFT-features are well known and frequently used in object recognition, for their rotation- and scale-invariance and performance in matching to other sets of features. The SIFT descriptors are represented as multi-dimensional vectors, abstracted from important points in an image, such as corners or edges. Once objects have been learned new objects are classified by finding the category of the most closely matching object in terms of SIFT. Objects are matched by finding the highest harmonic mean of two measures. In the first measure the number of visual features matched between the recognized and a learned object is divided by the number of features of the learned object, whereas in the second it is divided by the number of features of the recognized object. The category of the stored object with the highest score is picked as the name of the object recognized. For spatial relations the locations of objects are represented as average x, y and z coordinates of detected SIFT features.

[1] http://openkinect.org/wiki/Main_Page

3 Conversational strategies

The system learns objects either by being presented with them and told what they are (e.g. *This is a cup*) or by receiving feedback on an utterance it just made (*That's correct*). When the system hears a question such as *What is this?* (or a variation on this) it responds by also describing the certainty of its belief (*The object is thought to be a book, but it might also be a mug*). It can learn spatial relations when it recognizes both of the objects mentioned (*The book is on the right of the mug*). The system is also able to learn from feedback, confirmations of a human partner whether something was correct or not. The system may occasionally mishear the name of an object. The name can be unlearned right after learning (by saying *That is not what I said*), unlearned later (*Forget cup*) or re-learned to attach a new name to the previously learned object (*I said a book*). The system will occasionally ask the user for more examples of an object or spatial relation that it has too little knowledge of, but assumes the tutor takes the lead

	Accuracy	Accuracy cumulative
Round 1	96%	96%
Round 2	94%	95%
Round 3	96%	95.3%
Round 4	98%	96%

Table 1: Accuracy of recognition after the different testing rounds.

again right after that. This happens at random after a response or acknowledgement from the system.

4 Baseline evaluation

In the current experiment we test object recognition without human feedback. This will serve as a baseline for our forthcoming work where we will be testing incrementally more sophisticated interaction strategies that were described in the previous section. Ten objects are shown to the system for four rounds. After each presentation the system is queried for that object category. Note that although the object has not moved the system will make the classification from a new sensory scan. At each round the objects are placed in the same order and with approximately the same position and orientation.

5 Results and discussion

The accuracy of object recognition at each round as well as the cumulative accuracy over several rounds is presented in Table 1. These results show that accuracy of the system is very high and that it improves when more instances are learned. Table 2 shows the object matching scores over all object matches. The first column indicates objects presented to the system. The second column shows the average maximal matching scores (AMMS) with an object from the correct category (which may not be the winning one) over the four rounds, and the third column shows the corresponding standard deviations. High scores tell us that objects are easy recognisable, whereas low scores indicate that their recognition is more difficult. The fourth column shows the average overall matching scores (AOMS) against all object models, and the last column shows their standard deviations. This column demonstrates how much an object looks like any other object. Ideally, as we want objects to be uniquely distinguishable, AMMS should be high, while AOMS should be low.

Object	AMMS	Std. dev.	AOMS	Std. dev
Apple	.34	.07	.12	.10
Banana	.36	.07	.12	.10
Bear	.26	.06	.11	.06
Book	.50	.07	.19	.12
Cap	.15	.06	.10	.05
Car	.41	.06	.13	.11
Cup	.33	.10	.11	.09
Paint can	.22	.04	.11	.05
Shoe	.32	.01	.11	.08
Shoebox	.38	.07	.22	.11

Table 2: Object score and standard deviation.

6 Future work

In the immediate future we will examine the effects of varying object orientation and switching objects for other objects of the same category on the rate of learning. We will also test the learning of spatial relations. A change of interaction strategy will also be examined, starting with the contributions of feedback on learning and recognizing. An object ontology could also be implemented. The system could actively query users to gain information about how general the used term is, whether it is the name of a category or an object. As the learned databases are exportable, users could exchange these databases to increase the number of objects and spatial relations a system can recognize. Such a database could be made available on the internet, and divided into categories, depending on where the robot needs to work and what objects it will encounter. As the scale increases, however, it might become feasible to implement recognition with deep convolutional neural networks in favour of SIFT feature detection.

References

P. Lison. 2014. *Structured Probabilistic Modelling for Dialogue Management*. Ph.D. thesis, Univ. of Oslo.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1):170–205.