Eye Spy: Improving Vision through Dialog

Adam Vogel, Karthik Raghunathan, Dan Jurafsky
Stanford University
{av,kr}@cs.stanford.edu jurafsky@stanford.edu

Abstract

Despite efforts to build robust vision systems, robots in new environments inevitably encounter new objects. Traditional supervised learning requires gathering and annotating sample images in the environment, usually in the form of bounding boxes or segmentations. This training interface takes some experience to do correctly and is quite tedious. We report work in progress on a robotic dialog system to learn names and attributes of objects through spoken interaction with a human teacher. The robot and human play a variant of the children's games "I Spy" and "20 Questions". In our game, the human places objects of interest in front of the robot, then picks an object in her head. The robot asks a series of natural language questions about the target object, with the goal of pointing at the correct object while asking a minimum number of questions. The questions range from attributes such as color ("Is it red?") to category questions ("Is it a cup?"). The robot selects questions to ask based on an information gain criteria, seeking to minimize the entropy of the visual model given the answer to the question.

Introduction

Traditional methods for supervised training of vision systems require hand drawn bounding boxes or segmentations (Torralba, Murphy, and Freeman 2004). Although this is useful for expert designed systems, annotation is tedious and correctly segmenting objects takes some training.

Spoken language interaction is a natural interface to robots, requiring little training to use. Moreover, learning through a mixture of language and physical interaction is how children acquire language. Research on child language acquisition suggests that the negotiation of shared attention between the learner and language user is critical (Tomasello 2008).

We present ongoing work on a robotic system to learn names and attributes of objects through spoken interaction. Robot learning from human interaction has previously been applied to both perception (Roy et al. 2002) and manipulation (**?**). In our setting, the robot engages a human interlocutor in a spoken language game similar to "I Spy" and "20 Questions", popular games for child language learning. The robot uses feedback from this game to gather training examples for its vision system.

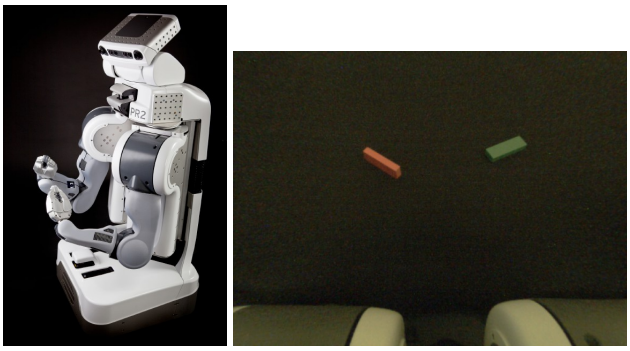

Figure 1: The Personal Robot 2 (PR2) and a sample scene from our dataset, captured with a camera on the PR2 head.

I Spy + 20 Questions

The robot interacts with a human interlocutor called Alice. They play a variation of "I Spy", where Alice picks out an object in a table top scene, and the robot tries to find it and name it. For instance, in Figure 1 Alice might choose the red block on the left of the scene, then say "I spy something". The robot then asks questions of both the object's category (e.g. "Is it a cup?") and attributes (e.g. "Is it green?"). At any point in the interaction the robot can guess the object location by pointing its arm, and Alice then tells the robot if it is correct. If the guess is right, the robot saves this example and all questions answered as a positive training instance, which is used to improve the vision system. Wrong guesses are saved as negative training examples.

Personal Robot 2

We use the Personal Robot 2 (PR2), produced by Willow Garage. Its sensors include a 3-DOF sensor head with wide-angle color stereo and a tilting laser range finder for accurate depth information.

Vision System

The vision system detects and localizes objects by *category* and additionally classifies their *attributes*. Category labels, drawn from a set C, correspond to concrete noun classes, such as "mug" or "stapler". Attributes correspond to noun

modifiers, such as "red", "wooden", or "rectangular", and we denote the set of attributes as A. Classifying object attributes allows us to better generalize to unseen objects - even if we don't know what something is, we might be able to tell that it is red and made of wood.

We use a sliding window object recognition architecture based on the STAIR Vision Library (Gould et al. 2010). Using standard image features based on color, shape, and texture we learn classifiers to predict both the category and attributes of an object. The output of the attribute classifiers is used as input to the category classifier, in a manner similar to (Farhadi et al. 2009).

In sliding window object recognition, we learn a probabilistic classifier to score the likelihood of an object of a particular category $c \in C$ with attributes $\vec{a} \in A$ occurring in a particular sub-window of the image b, denoted $p(c, \vec{a}|b)$. We then "slide" a window of varying sizes over all subregions of the image, and accept the most likely window as our guess. We use boosted decision trees for classification.

Dialog System

Our dialog system is powered by the Sphinx 3 Speech Recognition System paired with Festival speech synthesis. Speakers wear a Plantronics CS70 wireless head mounted microphone. Our current system asks yes or no questions, so speech recognition is easy.

Dialog researchers in psycholinguistics have modeled human question asking in terms of several metrics: information gain, KL divergence, impact, and others (Nelson 2005). We use an information gain heuristic to choose which question to ask next. Intuitively, we want to ask questions that most change our hypothesis about the location of the target object.

Each attribute $a \in A$ can be true or false, whereas an object has only one category $c \in C$. Questions $Q \in C \cup A$ are boolean, and ask the truth value of either an attribute a or a category membership c. Let b stand for the visual feature representation of a bounding box in the image. Furthermore, let $b^*(q)$ be the highest scoring bounding box under our current model, given the answer to question q,

$$b^*(q) = \arg\max_b \max_{c,\vec{a}} p(c, \vec{a}|q, b) \quad (1)$$

We want to ask a question $q \in C \cup A$ which maximizes the information gain of our classifier, which is equivalent to minimizing the conditional entropy of the category and attribute random variables for the most probable bounding box $b^*(q)$:

$$q^* = \arg\min_{q \in C \cup A} H(C, A|q, b^*(q)) \quad (2)$$

We can compute the information gain of a question q in terms of our vision model:

$$H(C,A|q,b^*(q)) = \sum_{(c,\vec{a}) \in C \times A} p(c, \vec{a}|q, b^*(q)) \log \frac{1}{p(c, \vec{a}|q, b^*(q))} \quad (3)$$

Pilot Study

We conducted a pilot study of the question asking strategy on a collection of four rectangular blocks of different colors. We collected a dataset of 100 images for each of the blocks and 300 images containing pairs of blocks. We solicited ground truth bounding boxes and color annotations from Amazon Mechanical Turk. For images containing two blocks, we then asked color attribute questions using our entropy minimization policy. After guessing the color of the target block, we then compare the vision model's most likely bounding box with the actual location of the block.

	Oracle	Random	Entropy
Average Questions	1	2.62	1.96

Table 1: Average number of questions asked per trial for different question asking strategies. The *Random* policy asks color attributes at random, *Oracle* always asks the correct question, and *Entropy* asks the question which minimizes the entropy of the resulting hypothesis.

Table 1 shows the average number of questions asked per trial and the accuracy the robot had in finding the correct block. These results show that the entropy minimization policy is effective at finding the correct object with few questions. Once our system knows which color block to find, it localizes the object with accuracy 94.06%.

Conclusion

Spoken dialog interfaces to robots have the potential for natural learning with humans. Evaluation of how examples gathered through the question asking interaction can improve object detection is ongoing work, but our initial results show promise. Additionally, we plan to apply our system to household and office items.

References

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. A. 2009. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Gould, S.; Russakovsky, O.; Goodfellow, I.; Baumstarck, P.; Ng, A. Y.; and Koller, D. 2010. The STAIR Vision Library (v2.4). http://ai.stanford.edu/ sgould/svl.

Nelson, J. D. 2005. Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological Review* 112(4):979–999.

Roy, D.; Gorniak, P.; Mukherjee, N.; and Juster, J. 2002. A trainable spoken language understanding system for visual object selection. In *In Proceedings of the International Conference of Spoken Language Processing*.

Sorokin, A., and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops*.

Tomasello, M. 2008. *The Origins of Human Communication*. MIT Press.

Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2004. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 762–769.