Structural Descriptions In Human-Assisted Robot Visual Learning

Geert-Jan M. Kruijff & John D. Kelleher
LT Lab
DFKI GmbH
Saarbrücken, Germany
{gj,kelleher}@dfki.de

Gregor Berginc &
Aleš Leonardis
ViCoS Lab
University of Ljubljana
Ljubljana, Slovenia
{gregor.berginc,alesl}@fri.uni-lj.si

ABSTRACT

The paper presents an approach to using structural descriptions, obtained through a human-robot tutoring dialogue, as labels for the visual object models a robot learns. The paper shows how structural descriptions enable relating models for different aspects of one and the same object, and how being able to relate descriptions for visual models and discourse referents enables incremental updating of model descriptions through dialogue (either robot- or human-initiated). The approach has been implemented in an integrated architecture for human-assisted robot visual learning.

Categories and Subject Descriptors

I.2.7 [**AI**]: Natural language interfaces; I.2.10 [**AI**]: Vision and Scene Understanding

General Terms

Algorithms

Keywords

Cognitive vision and learning; natural language dialogue

1. INTRODUCTION

One meaningful dimension of human-robot interaction is the ability for a robot to connect vision and language. A crucial problem therefore is *perceptual grounding*. Approaches to perceptual grounding focus on how to learn models that connect words or sequences (i.e. *expressions*) thereof to perceptual features; cf. [4] and references therein. This yields a grounded representation that provides a level of perceptual understanding which purely symbolic object descriptions traditionally lack.

Most authors refer to this perceptual understanding as the meaning of an expression. This is true only insofar as we just consider that meaning in isolation. If we want to understand the meaning of an expression used in the context of a dialogue, the representations we assign should also enable *linguistic grounding* besides perceptual grounding. Representing an expression as a string does not provide enough structure for this – we lack e.g. the means to relate the occurrence of an expression to the preceding dialogue context.

We propose to enhance the characterization of a visual model of an expression with a *structural description* of its linguistic meaning, seen as an *ontologically rich relational structure*. These descriptions enable us to reflect the use of an expression, and its visual reference, in a dialogue. This way, we can incrementally update or learn the description of a visual referent. By co-indexing descriptions we can also explicitly identify models for different aspects of a specific (type of) object. Finally, structural descriptions give the linguistically expressible properties of a visual model for the perceptual meaning of an expression, so they do not replace models for grounding but complement them, e.g. [4].

Below we discuss structural descriptions and their use, e.g. how identification across models and incremental updating are handled. We briefly present the implementation of this approach in an integrated architecture for human-assisted robot visual learning.

2. STRUCTURAL DESCRIPTIONS

(1) gives a simple, tutor-driven dialogue. The tutor fully describes an object, the robot acknowledges it has understood. The robot labels the model it learns (Figure 1(l.)) with the structural description obtained from the analysis of the tutor's utterance (2). The logical description in (2) states $b2$ as the identifier, of sort *thing*, being a box with a property of having a color orange; [2].

(1) H.1 "This is an orange box."
R.2 "Okay."

(2) $@b2 : thing(\mathbf{box} \wedge \langle Property \rangle (o1 : color \wedge \mathbf{orange}))$

We can handle incremental updating of structural descriptions by relating identifiers for discourse referents to the identifiers for structural descriptions of visual object models. In (3), the tutor first (H.1) provides only a partial structural description (4a), and only later (H.2) completes it with the addition of a property ascription (4b). Discourse analysis resolves the pronoun "It" (H.2) to refer to the box, i.e. the property ascription "It is orange" applies to the box talked about earlier ($t1 = b2$, yielding the description in (2)).

(3) H.1 "This is a box."
H.2 "It is orange."

HRI'06, March 2–4, 2006, Salt Lake City, Utah, USA.
ACM 1-59593-294-1/06/0003.

R.3 "Okay."

(4) a. $@b2 : thing(\textbf{box})$

b. $@t1 : thing(\langle Property \rangle(o1 : color \wedge \textbf{orange})) \wedge$ t1=b2 $\Rightarrow$ (2)

(5) below provides an alternative to (3). In (3), the incremental update of the description for the object model was tutor-driven. In (5) the robot prompts the tutor for more information, asking a *wh*-question.

(5) H.1 "This is a box."
R.2 "Okay."
R.3 "What color is the box?"
H.4 "It is orange."

(5) assumes the robot can establish whether an object description is complete. We exploit here the "ontologically promiscuous" nature of the representations to assess descriptive completeness, by using ontologies for object types and their associated properties. Connecting structural descriptions to object ontologies through the object type also enables us to check for inconsistencies in a description.

Using co-indexation we can not only relate structural descriptions to discourse referents, but also to other structural descriptions. If the tutor follows up the dialogue in (4) with (H.4) "This is its side", the robot acquires a model (Figure 1(r.)) with a structural description that we can link to the model described in (2) by reusing the identifier $b2$ after resolving the antecedent for "its": $@s1 : thing(\textbf{side} \wedge \langle Partitive \rangle(b2 : thing \wedge \textbf{box})$.

Figure 1: Front (l.) and side (r.) of orange box $b2$

3. IMPLEMENTATION

We have implemented the approach in a distributed architecture for integrating different perceptual and deliberative skills that deal with a variety of modalities. The architecture is inspired by multi-level distributed cognitive architectures.

The communication subsystem consists of several components for the analysis and production of natural language. It has been implemented as a distributed architecture using the Open Agent Architecture[1]. Analysis starts with Sphinx4 speech recognition[2]. The string-based output of Sphinx4 is parsed with OpenCCG[3]. OpenCCG employs a combinatory categorial grammar to yield a representation of the linguistic meaning that the string (i.e. the utterance) represents [2]. We represent linguistic meaning in the same description logic-like formalism we use for structural descriptions for visual object models. Finally, in dialogue analysis we relate the linguistic meaning of an utterance to the current dialogue context, in terms of how it rhetorically and referentially relates to preceding utterances. This yields an updated model of the (situated) dialogue context [1]. On the production side, we use dialogue planning to enable flexible, contextually appropriate interaction. Based on a need to communicate, established either by the current dialogue flow or by another modality, the dialogue planner establishes a communicative goal. In turn, we plan the content to express this communicative goal, possibly in a multi-modal way. In these planning steps, we can inquire the models of the situated context (e.g. dialogue context, visually situated context) to ensure that the content we plan is contextually appropriate. We realize verbal content using the OpenCCG realizer, which generates a string for the utterance, and then synthesize this string using a text-to-speech engine[4].

In the vision subsystem, we have implemented visual scene understanding based on three cues: identity, color, and size of objects in the scene. We use SIFT (Scale Invariant Feature Transform) features [3] to recognize object identity - shown as the white circles in Figure 1. Each SIFT feature is a vector $(x, y, \theta, \sigma, v)$, where $(x,\ y)$ gives the position of the feature, θ the main orientation and σ the scale at which the feature was detected. We store the description of the local patch around $(x,\ y)$ as a 128-dimensional vector v.

To reason about colors, the robot estimates the bounding box of the object. When recognizing, the robot detects features, and tries to match them with the features stored when learning the model. If the number of matches is over a given threshold, the affine transformation is estimated based on affinities between matched features. We obtain the pose of the object by applying this affine transformation to the model's segmentation mask. The robot calculates the color histogram over the segmented region of the IHS color space. The peak of smoothed histogram indicates the color.

Each time the tutor initiates learning, by saying e.g. "This is an $< X >$", the robot collects SIFT features and labels them with the structural description of X. For training, we currently assume the scene contains only a single object, which the tutor is talking about. If the robot already knows the object X, it tries to update its representation. To improve robustness, the robot uses several consecutive frames and uses only features that remain stable.

The vision subsystem consists of several CORBA[5] servers. We use an OAA agent to serve as a mediator between the communication subsystem and the vision subsystem.

4. ACKNOWLEDGMENTS

The research reported in this paper was supported by the EU FP6 IST Cognitive Systems Integrated project *Cognitive Systems for Cognitive Assistants* "CoSy" FP6-004250-IP.

5. REFERENCES

[1] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.
[2] J. Baldridge and G.J.M. Kruijff. Coupling CCG and hybrid logic dependency semantics. In *Proc. ACL 2002*, Philadelphia, PA, 2002.
[3] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *Int. Jnl. Computer Vision*, pages 91–110, 2004.
[4] D.K. Roy. Semiotic schemas: A framework for grounding language in the action and perception. *Artificial Intelligence*, to appear.

[1] http://www.ai.sri.com/oaa/
[2] http://cmusphinx.sourceforge.net/sphinx4/
[3] http://openccg.sf.net
[4] http://mary.dfki.de
[5] http://www.corba.org/