BISHOP|BLENDER: Spatially Grounded Language Understanding in 3D Modelling Software

Peter Gorniak
MIT Media Laboratory
20 Ames St.
02139 Cambridge MA
pgorniak@media.mit.edu

Deb Roy
MIT Media Laboratory
20 Ames St.
02139 Cambridge MA
dkroy@media.mit.edu

We present BISHOP|BLENDER, an augmentation of the 3D modelling application Blender (Blender Foundation , 2003). In the BISHOP project (Gorniak and Roy, 2004) we investigated how people describe objects in visual scenes using spatial language, and built a visually grounded language understanding system that performed well in understanding visually referring expressions. One application of this work is in applications that share a virtual world with the user, such as 3D modelling applications. Specifically, we here address the problem of selecting objects in a complex and cluttered 3D scene such as that shown in Figure 1. Using a 2D pointing device such as a mouse to select objects in a 3D scene is error-prone, because many objects at different distances from the viewer may share the same 2D screen space. Users currently have to solve this problem by manually rotating the scene to isolate the intended target in its own 2D area (as task that can be hard in a cluttered scene), or by using a 'select other' function that targets an object at a different depth in the same screen area. Both methods are cumbersome at best. As an alternative, we demonstrate the use of a language understanding system like BISHOP to resolve referents in the 3D scene. In BISHOP|BLENDER the user can use spoken commands like "Select the door behind that" or "View the leftmost window" to select objects in absolute terms or relative to the current selection, while continuing to work with the modelling application as usual.

The BISHOP|BLENDER project extends the BISHOP system in a number of ways:

- Instead of constrained virtual scenes, BISHOP|BLENDER works with arbitrarily complex 3D models in a real 3D modelling application as seen in Figure 1
- BISHOP|BLENDER grounds language in the 3D configuration of the scene, as opposed to the 2D projection of the scene that BISHOP used. This allows BISHOP|BLENDER to understand a full set of spatial relations, including "above","to the left of" and "behind".
- BISHOP|BLENDER understands language relative to an arbitrary viewpoint in 3D space, as opposed to the restricted frontal viewpoint used in BISHOP.
- whereas experiments to evaluate BISHOP were performed on transcribed speech, BISHOP|BLENDER uses the 4 pure Java speech recognizer to directly understand the user's speech commands (Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Laboratories , 2004).
- BISHOP used different methods to visually ground spatial relations ("to the left of"), spatial extrema ("leftmost"), and colour terms ("green"). In BISHOP|BLENDER we have unified these groundings into a single algorithm based on Tenenbaum's word generalization paradigm (Tenenbaum and Xu, 2000). The advantage of this example based approach is that very few examples are required. We

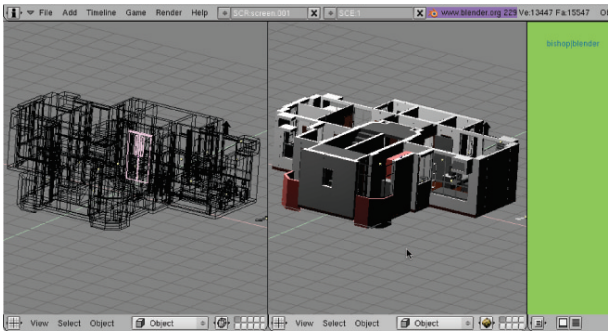

Figure 1: A typical blender window showing a model of an apartment in a shaded and a wireframe version, with an interior door selected. The user can now use speech to select another object, for example by saying "select the one behind that one"

extend the approach linearly interpolate between what we call background and foreground examples. Background examples are known independently of the current scene, such as patches of blue that were labelled "blue" before. Foreground examples are in terms of the relevant property values (such as RGB colour or view-relative spatial location) normalized to lie between 0 and 1. We apply Tenenbaum's generalization algorithm twice, once with the absolute property values of the currently considered objects using the background examples, and once with the property values of the currently considered objects normalized to lie in between 0 and 1 using the foreground examples. A single weight interpolates in between the two resulting estimates. If this weight is set to favour background examples, it achieves generalization behaviour appropriate to, for example, colour terms, where the scene can be divided into objects that are green and those that are not. If the weight is set to favour foreground examples, it captures behaviour appropriate to spatial location terms, where there is always at least one "leftmost" object if there are any objects at all.

- We have added some rudimentary verb commands to blender to distinguish between selection (using "select" or an object description without a verb) and changing the viewpoint to focus on one or more objects ("view the doors").

In the future, we would like to merge the work represented in the BISHOP|BLENDER project with the speech learning interfaces from some of our other work (Gorniak and Roy, 2003). Currently, the user must specify the names of objects (in the examples in this paper "door" and "window") by typing them in at object creation time. Applying work that allows users to train speech commands for interface events to BISHOP|BLENDER would allow us to learn phonetic names for objects in the scene and use these instead of manually provided strings to let the user refer to objects in the scene by name.

References

Blender Foundation. 2003. Blender 3D graphics creation suite. http://www.blender3d.org.

Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Laboratories. 2004. Sphinx 4 Java Speech Recognizer. http://www.speech.cs.cmu.edu/cgi-bin/cmusphinx/twiki/view/Sphinx4/WebHome.

Peter Gorniak and Deb Roy. 2003. Augmenting user interfaces with adaptive speech commands. In *Proceedings of the International Conference for Multimodal Interfaces*.

Peter J. Gorniak and Deb Roy. 2004. Grounded compositional semantics for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Josh B. Tenenbaum and Fei Xu. 2000. Word learning as bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.