# *The Blue One to the Left*: Enabling Expressive User Interaction in a Multimodal Interface for Object Selection in Virtual 3D Environments

Pulkit Budhiraja and Sriganesh Madhvanath
Hewlett-Packard Labs, India
24 Salarpuria Arena, Hosur Main Road, Adugodi
Bangalore 560030, INDIA
+91 80 33829169
pulkit.budhiraja@gmail.com, srig@hp.com

## ABSTRACT

Interaction with virtual 3D environments comes with a host of challenges. For instance, because 3D objects tend to occlude one another, performing object selection by pointing gestures is problematic, and more so when there are many objects in the scene. In the real world we tend to use speech to clarify our intent, by referring to distinctive attributes of the object and/or its absolute or relative location in space. Multimodal interactive systems involving speech and gesture have generally relied on speech for commands and deictic gestures for indicating the target object. In this paper, we present a system which allows object references to be made using gestures and speech, and supports a variety of expressions inspired by real-world usage.

## Categories and Subject Descriptors

H.5.2 **[Information Interfaces and Presentation]:** User Interfaces – *Interaction Styles.*

## General Terms

Design, Human Factors

## Keywords

Multimodal interface, selection, virtual 3D environment

## 1. INTRODUCTION

Multimodal interactive systems involving speech and gesture have generally relied on speech for commands and deictic hand or touch gestures for indicating the target object. The use of gestures for pointing becomes problematic when there are a large number of objects close together, and gets worse when applied to virtual 3D environments wherein 3D objects tend to occlude one another. In the real world we tend to use speech to clarify our intent, by referring to distinctive attributes of the object (e.g. "*the big blue book*") and/or its absolute or relative location in space (e.g "*the one in the left corner*").

In this paper, we present a system which allows object references to be made using a variety of expressions that we use to refer to objects in the real world, for instance: (i) using speech and gesture in combination, (ii) using visible attributes, of objects, (iii) indirect references involving spatial relations, (iv) references to spatial regions, and (v) anaphoric references.

The intent of supporting multiple ways of referring to objects is to give the user freedom to use the most convenient option to refer to an object in a particular scenario, rather than being constrained to follow a specific command syntax or interaction technique.

*ICMI '12*, October 22–26, 2012, Santa Monica, California, USA.
ACM 978-1-4503-1467-1/12/10.

## 2. RELATED WORK

Related work in the field of multimodal object reference includes Peter Gorniak's *Bishop*, which understands natural language references to objects in a spatial scene [5]. Edward Tse's *speech-filtered bubble ray* uses speech as a filter for selecting objects in a densely packed area on a wall display using pointing gestures [1]. Landragin's work focuses on classifying objects under multiple reference domains based on their visual perception, and gestures and language used in a dialogue [3]. Our system is greatly inspired by work of Makalic and Zuckerman, which focuses on probabilistic reference disambiguation for spoken dialogue [2][4].

## 3. INTERFACE AND INTERACTIONS

The physical setup of the system includes a large display attached to a Windows PC (Fig.1). A 3D scene is rendered by computer graphics software called *Blender*. Hand gestures are captured using a *Microsoft Kinect* depth camera. Speech is captured using a headset microphone. The user stands at a distance of 4-6" from the display to interact with the system.

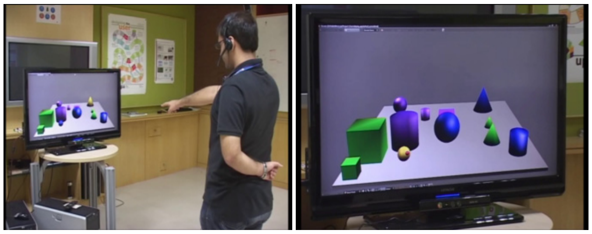

**Figure 1. Physical setup and 3D scene**

The 3D scene in the present version comprises of simple objects such as cubes, cylinders, cones and spheres which may differ in size and color. The present version only allows selection of objects, and of one object at a time. However the design is extensible in order to allow selection of groups of objects, as well as other commands such as deletion, movement and so on.

The system supports object selection through speech using restricted natural language and/or gestures. The system provides the user with multiple ways of interacting with the scene. References to objects may be made using speech or both speech and gesture. The system also attempts to understand grammatically incorrect utterances. The following types of object references can be understood by the system:

***Object Attributes*** The user can refer to an object using one or more visible attributes such as color, shape, size, e.g. "*the big blue cone*" as well as its spatial location (by pointing). The user has the freedom to choose as many attributes as he/she desires to specify the object. The system uses these attributes to obtain an unambiguous reference to the target object.

***Spatial Regions*** The number of objects to be considered can be filtered by making a reference about the region of the target object, e.g. "*the blue one to the left*". The system compares an object's position in the scene with a reference point depending upon the region specified. For example, for "*left*" the leftmost point of the scene is taken as the reference point. The system supports references to *left*, *right*, *front*, *back*, *center* and *corner*.

***Spatial Relations*** For objects that are difficult to refer to directly, an indirect reference with respect to another object can be made, e.g. "*the object behind the green cone*". Indirect references can also be with respect to the currently selected object, e.g. "*the blue cone behind that one*". The system can resolve the different meanings of reference depending upon the context of the utterance. The system supports a variety of spatial relations between the two objects, such as *left*, *right*, *front*, *back*, *behind*, *ahead*, *near*, *next*.

The system also allows the user to rotate the scene using speech or gesture commands. A video demonstration of the system can be seen at http://vimeo.com/44531259.

## 4. SYSTEM DESIGN

Abstract information about all the objects in the scene is known to the system (Fig. 2). The *User Display Application* (*Blender*) renders the scene using this information. In order to interpret interactions, the system analyzes multiple transcripts of the speech and employs a probabilistic semantic engine to assess the meaning of spoken utterance. The probabilistic approach of understanding an utterance enables the system to handle ambiguity and speech recognition errors.

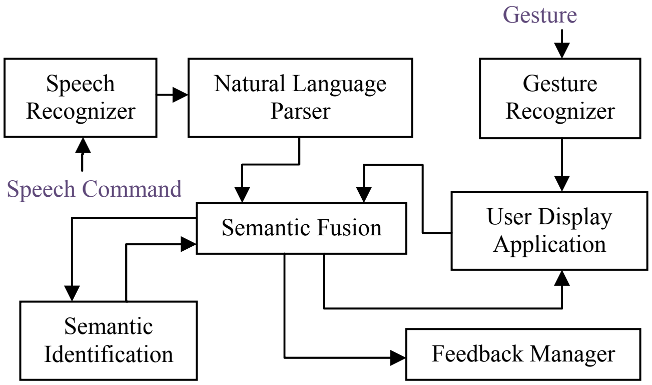

**Figure 2. System design**

The *Speech Recognizer* (*Microsoft SAPI)* returns multiple interpretations of the spoken utterance, each with its confidence score. The *Natural Language Parser* (*Stanford NLP*) returns a parse tree for each interpretation. The *Gesture Recognizer* sends the screen coordinates of the user's pointing hand to the User Display Application. These coordinates are mapped to a pointer in the scene for visual feedback. The parse trees and hand coordinates are sent to the *Semantic Fusion* component.

***Semantic Fusion*** Semantic Fusion encompasses temporal and semantic fusion of speech and pointing gestures. On-screen locations of objects referred to by pointing are determined by averaging the hand coordinates corresponding to spoken keywords such as "*this*" or "*that*" in the utterance.

The verb phrase, noun phrases and prepositional phrases in the parse tree are examined to extract information about the action, attributes of the target object and (in case of indirect reference) referred object and the spatial relation between them. This information is represented in a graph-like data structure. Multiple such graphs are generated corresponding to alternative parse trees, and sent to the *Semantic Identification* component to resolve object references in the interaction, to objects present in the scene.

***Semantic Identification*** The Semantic Identification component tries to match the attributes of each of the objects in the scene, e.g. color, shape, size and location, with those extracted from the parse tree, and returns a likelihood score. If the target object is to be restricted to a particular region, the identification layer assigns a score to a candidate object by comparing a specific feature dimension (depending upon the region specified) with the absolute dimensions of the scene using a Gaussian function (Fig. 3). In the case of indirect references, the spatial relation between the candidate target object and the candidate referred object is evaluated. The target object from the graph having the best combined likelihood score is selected as the result.

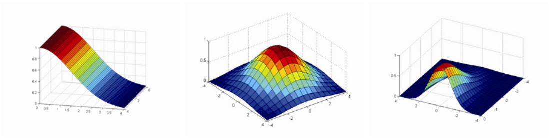

**Figure 3. From left to right: Probability distributions for spatial regions *left* and *center*, and spatial relation *behind***

## 5. SUMMARY AND NEXT STEPS

In this paper, we have described a prototype system for interacting with virtual 3D environments that supports expressive object references made using gestures and speech. The system is presently limited in the types of objects that can be present in the scene, and the attributes that may be used to refer to them (shape, color and size). These restrictions would need to be addressed for creating useful real-world applications. In terms of possible actions beyond selection; movement, rotation, deletion, etc. can be added to the system, requiring the semantic fusion and semantic identification logic to be suitably extended. The user interface of the application can be further improved by adding the ability to zoom and pan using speech, gestures or both.

## 6. REFERENCES

[1] Edward Tse, Mark Hancock, and Saul Greenberg. 2007. Speech-filtered bubble ray: improving target acquisition on display walls. In *Proc 9th Intl Conf on Multimodal Interfaces* (ICMI '07)

[2] Enes Makalic, Ingrid Zukerman, Michael Niemann, and Daniel Schmidt. 2008. A Probabilistic Model for Understanding Composite Spoken Descriptions. In *Proc 10th Pacific Rim Intl Conf on AI* (PRICAI '08)

[3] Frédéric Landragin. 2006. Visual perception, language and gesture: a model for their understanding in multimodal dialogue systems. *Signal Processing* 86, 12

[4] Ingrid Zukerman, Enes Makalic, Michael Niemann, and Sarah George. 2008. A Probabilistic Approach to the Interpretation of Spoken Utterances. In *Proc 10th Pacific Rim Intl Conf on AI* (PRICAI '08)

[5] Peter Gorniak and Deb Roy. 2003. A visually grounded natural language interface for reference to spatial scenes. In *Proc 5th Intl Conf on Multimodal interfaces* (ICMI '03)