Exploration in a Large Corpus: Research on the Integration of Eye Gaze and Speech with Visual Information in a Virtual Reality System

C. R. Voss and **J. Gurney**
Natural Language Processing Group
Army Research Laboratory
Adelphi, MD

J. Walrath
Computer Visualization Branch
Army Research Laboratory
Adelphi, MD

Introduction

With the availability of large TIVA (text, image, video, audio) corpora in digital form has come the call for "effective" computational methods to access the content of these corpora. Our research has focussed specifically on *interactive* access to *one class* of corpora, the large terrain and object data sets that make up a virtual reality (VR) world.

We are currently experimenting with a new interactive computational method to assist a person in exploring the content of our VR system: by tracking their eye gaze for points of fixation on the VR screen while they are talking about what they see and what they want to see, we can analyze how people integrate *visual searching* and *verbal searching* during real-time exploration of images and changing scenes. In the first phase of this research we ask if a VR user's gaze points can disambiguate the referents in the natural language (NL) speech they produce and thereby increase the accuracy of their access to the VR world objects and data.

Background

Our research approach to understanding interactive exploration of the content of a virtual world, in particular, and a large corpus in general is empirical:

- Are there observable, recordable behaviors[1] that people readily produce in different modalities while searching through a large corpus that indicate what is being attended to?
- Are there regular co-occurrences in such behaviors that could be used to guide an interactive computational search of other corpora?

The second question addresses the issue of the generality of this method beyond a VR system, whether it could also yield a way of assisting a person to explore or navigate through a "space" with a wide variety of content, such as within a digital library with a mix of TIVA documents.

With our application, the various sources of data that constitute the corpus are merged as needed and rendered at run-time onto a screen for display to the user. The user can be thought of as peering into the corpus through a screen-size window and seeing a continuous presentation of the data sources already merged. Although this merged presentation differs with the discrete document or image presentations found in searches through collections of media, the technique we are experimenting with could, in principle, be applied to navigating in these corpora as well.

This work takes place in the context of our larger ongoing research effort to understand—for interactive access to entities, relations, and operations in a VR world—which computational methods will properly interpret:

1. how people use natural language to explore the very large terrain and object image corpus, e.g., how do they talk about objects and their spatial relations?
2. how people naturally use their eyes to explore a scene as they hear it being described, e.g., what are the patterns of gaze fixation?
3. how these two modalities—speech and eye gaze—are integrated in the exploration, e.g., how do the objects being referred to in a person's speech show up as points of attention in their visual field?

We briefly describe the VR system we are using as well as our recent research on "natural language navigation" in that system, where the VR users' natural language expressions will locate and relocate objects (including their own point of view) on the VR screen. We then explain the limitations of this NL-only navigation and motivate our current research on the collection and analysis of concurrently recorded eye gaze, speech and scene (EGSS) information.

Visual Information in a Virtual Reality System

The Army Research Laboratory has designed a real-time visualization system that is a version of the Virtual GIS (VGIS) (Kesslet 1995) that was built using the Simple Virtual Environment Library (SVE) (Koller draft). This system displays terrain models derived from elevation and imagery data, along with GIS

[1] Our research currently examines speech and visual gaze.

raster layers, protruding features, buildings, vehicles, and other objects.

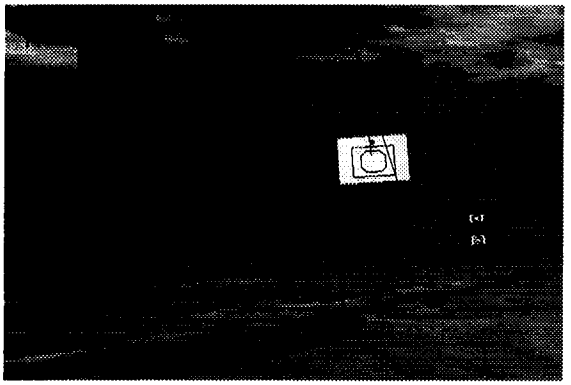

Figure 1: Scene in Virtual Reality System

Users run interactive scenarios in real-time. The observation point-of-view, i.e., the perspective of the user, can be attached to the terrain or repositioned by the user. It can also be attached to a moving (land or airborne) vehicle in the scenario or a vehicle controlled by the user. Both the window-based (perspective 3-D) and immersive (stereo 3-D) versions of the VR system allow flight through six degrees of freedom; to aid navigation, there is a pop-up overview map co-ordinated with the scene.

A typical VR scene appears in figure 1. This sort of view presents a lot of information of various kinds: there is the picture itself which represents part of the world as it would appear from the chosen viewpoint; there is the imposed symbology (the "flags" and connecting blue lines) which represent the organizational groupings and hierarchy of the various vehicles, platoons, companies, etc.; and there are the vehicles and other objects which can be queried for several types of information. In principle it is possible to ask questions about any of the objects in this domain – both those visible in the scene and those not currently in the field of view. Thus the VR scene provides a window into the data base.

Our research is focussed on how a user who is looking at this scene can use NL to extract information from the database.[2] We are currently building a NL navigation tool for the VR system (Gurney, Klipple, & Voss 1996). The terrain and object data[3] that form the corpus is loaded as the initial VR world during the startup phase (see figure 2).

[2]This VR system was not originally designed for and no allowances were made for, a speech-based natural language interface. All user interaction was intended to be channeled through a set of pull-down menus and a mouse pointer.

[3]The objects are stored with numeric and text data to encode shape, name, location, and other properties.

The Problem: Questions that Occur Relative to Visual Presentation

This is a bimodal system: *spoken* NL queries will be about objects *visible* in the scene. In this way the scene serves as a window into the data. To take full advantage of this window, we must now provide a robust solution to the question: How will users talk about what they see in the scene? In particular, how can we disambiguate NL questions that are essentially about the visible objects and relations? This is a new problem for query processing because the interpretation of a query will now depend on what the user is looking at (as we will illustrate below).

NL Navigation

We define *natural language navigation* as the task of using expressions of natural language to locate and relocate objects—including one's own self—within some real or virtually imaged physical space. Expressions in NL navigation such as:

Find the tank behind the building.
Drive the tank between the buildings.

require identifying the referring objects (here, tanks and buildings) and then selecting the appropriate operation with respect to these objects.

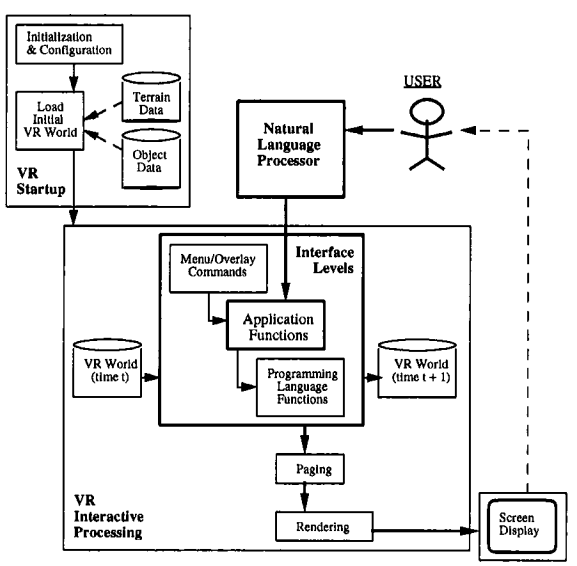

Figure 2: The VR Portion of the System

Limitations of NL-only Navigation

One of the difficulties we encounter is the need to disambiguate the *linguistic* referent in a spatial expression in terms of the available set of *rendered* and *stored* objects in the virtual world, corresponding to images seen on a screen and entities unseen but accessible in the VR

world's data base. Given the scene shown in figure 1, the following questions would be appropriate:

Where is the rest of that platoon?
What is the strength of that first tank?

The noun phrase *that platoon* refers to the platoon indicated by the large symbol floating above the three foreground tanks. Therefore, our NL query processor needs to detect that this is the symbol of interest to the user and then look up the associated platoon.[4] The problem in this case is that, in the distant background, there is actually one other platoon symbol. This one is quite small and, therefore, probably not the one of interest to the user (it may have gone unrecognized). So, one kind of processing would reason about near objects and far objects in order to find the intended object. A more direct and perhaps more robust method would be to assume the user's point of attention by tracking the user's point of gaze.[5] This is the approach we will motivate and propose in the next section.

The second example above, refers to a *first* tank. Presumably, given the VR scene, this could be relative to some spatial ordering. In this case there are three tanks motoring down a road. On one ordering the first tank is the one leading the convoy. But on another ordering the closest tank is the first one in order of receding distance from the user. Again, the attentional state of the user should be relevant to the required disambiguation. We observe that these and other dependencies on the user's visual attention will occur precisely because a visual scene is affecting the user's use of language. In the next section we discuss some previous work on the use of eye gaze tracking to uncover the attentional states of human subjects. We then discuss our plan to investigate the application of eye gaze tracking to our NL query processing problems.

Working Solution

So far we have described our VR system and its corpus, the NL component that enables a system user to navigate through the VR world "hands-free" using only NL expressions, and the problem of object reference resolution when only natural language is available for exploring within the VR system. Here we present our working solution to this problem, augmenting the NLVR system by non-intrusively monitoring the eye gaze points of the user.

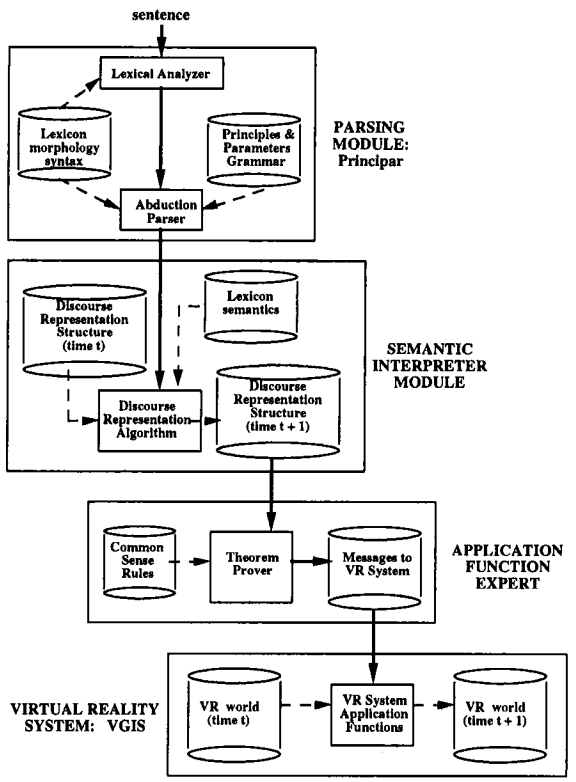

Figure 3: The NL Portion of the System

Eye Gaze Information

Psycholinguists have recently developed a new research paradigm to test the real-time comprehension of spoken language commands: they track the *eye gaze* of their subjects throughout the time that the subjects hear and physically respond to these commands (Tanenhaus *et al.* 1995). These experiments demonstrate that it is feasible to monitor the ongoing process of understanding natural language by way of the eye movements of the subjects involved. For example, they found that when asked to touch one of several blocks presented, a person made an eye movement to the target block an average of 250 ms after the end of the word uniquely specifying that target.

Tanenhaus *et al.* (1995) also found that, under certain conditions, what the subjects saw enabled them to respond within 145 ms to what they were hearing. That is, their visual context had an effect on the speed of their response to NL input. Or, put another way, the two channels of input, the auditory linguistic input and the visual input, were not independent.

This work provides the starting point for our research. Although this work examined linguistic and visual processing only during *comprehension* tasks, it

[4] Other processing is required to detect which are the *other* platoon members. The other members are the ones not visible in this scene.

[5] Several researchers have helped to establish the corrolation between the empirical *point of gaze* and the cognitive *point of attention* (Yarbus 1967), (Kahneman 1973), (Klein 1980), (Rizzolatti *et al.* 1987). Also note that these terms are distinct from *viewpoint* (or *point of view*) that refers to the property of the VR system display presented to the user. For example, a display may have an *inside-to-outside* or an *outside-to-inside* viewpoint (Kantowitz & Sorkin 1983).

suggests that these two processing modalities may also not be independent during *production* tasks either. That is, a person's speech or linguistic output and that person's gaze point may not just be coincidentally parallel signaling systems, but rather related in linguistically relevant ways. At this point, what we do not yet know on the basis of empirical data—and what is relevant to the linguistic problem of reference resolution described above—is whether a person's real-time visual exploration in a VR system is related to their own production of spoken language expressions.

Focus of Attention

It is also well known that focus of attention helps determine the referents of ambiguous noun phrases in non-visual, discourse situations. The phenomenon of "centering" has been proposed by (Grosz, Joshi, & Weinstein 1996), by (Brennan, Walker, & Pollard 1987), and by others to account for discourse anaphora. We interpret these results as further evidence that the intentions of a speaker can be determined by what they are attending to. In the case of discourse, the attended objects can be uncovered from the structure of the discourse. We conjecture that these objects can also be uncovered by detecting the subjects history of visual attention.

Augmented Navigation

We are augmenting the NLVR system with an eye tracker that can measure, record and analyze where a system user in looking on the VR display (see (Lahoud & Cleveland 1994) for details of the Eyegaze Development System we are using).[6] While this particular tracker may also be used in an HCI command/control mode (where users choose to actively direct their eyes as if pointing with them), this is *not* the usage in our system. For our application, the tracker provides a non-intrusive "passive monitoring" mechanism for finding the users' gaze point. No attachments to a user's head are required.

Image processing algorithms in the tracker's software locate the relative positions of the pupil center and the corneal reflection in order to compute the user's gaze point on the screen. These measures are computed from the image received via a video camera that continually observes the subject's eye. For each user, a short (under 1 minute) one-time calibration procedure is needed to determine their specific relevant measurements (the cornea's radius of curvature, angular offset between their eye's optic an focal axes).

The application software computes raw eyegaze data each 30th or 60th of a second and stores this for analysis of the eye's saccades and fixation points. The locations of the fixation points are then available for analysis within the NLVR system for identifying which object in the scene is being looked at. Recall that in the object reference resolution problem, the user's NL referring expression may be ambiguous; there may be more than one object in the scene that the user could be talking about. The empirical question for our research now is to determine if gaze points will help identify the user's intended referent.

[6] See (Young & Sheena 1975) references to other methods of measuring eye orientation.

Research Questions

In order to increase our understanding of how different media may be more effectively understood and searched computationally, it behooves us to do the empirical research to collect observable data on what people readily attend to as they do such tasks, determining what they are saying and looking at when they explore a "space". Thus we have begun to look specifically at observable behaviours in order to ask how people integrate their *visual searching* (by eye gaze) and *verbal searching* (by speech) during real-time exploration of images and changing scenes.

Our ongoing research now involves developing the experiments to collect this observable data with the aid of the eye tracking mechanism. The goal of our experiments is to determine whether a person's eye gaze and speech are integrated in a way that is linguistically significant while they navigate through a virtual world. More specifically, do the patterns of gaze points in the user's visual search help disambiguate linguistic referring expressions in the user's verbal search? The experiment requires calibrating the subjects' eyes, establishing the subjects' gaze points appropriately for the eye tracker's video monitor, and then eliciting natural speech from the subjects that includes: (i) sentences with unambiguous linguistic referents in the scene, (ii) sentences with no linguistic referents in the scene, and (iii) sentences with an ambiguous linguistic referent in the scene.

If a well-correlated relation between the linguistic referents (in the verbal search) and the gaze points (in the visual search) is found in the experimental data, then the accuracy with which the content of VR scene can be identified and navigated will increase. In principle then, for each well-correlated or predictive relation that is discovered, the user of the augmented NLVR system will have a more accurate basis for exploring the VR world.

References

Brennan, S. E.; Walker, F. M.; and Pollard, C. J. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the ACL.*

Grosz, B. J.; Joshi, A. K.; and Weinstein, S. 1996. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics* 21:203–226.

Gurney, J.; Klipple, E.; and Voss, C. 1996. Talking about what we think we see: Natural language pro-

cessing for a real-time virtual environment. In *Proceedings of the IEEE Symposium on Image, Speech, and Natural Language Processing.*

Kahneman, D. 1973. *Attention and Effort.* New Jersey, NJ: Prentice-Hall.

Kantowitz, B. H., and Sorkin, R. D. 1983. *Human Factors: Understanding People–System Relationships.* New York: John Wiley & Sons.

Kesslet, D. 1995. *The Simple Virtual Environment Library.* Atlanta, GA: Georgia Institute of Technology.

Klein, R. 1980. Does oculomotor readiness mediate cognitive control of visual attention? In Nickerson, R. S., ed., *"Attention and Performance VIII".* Hillsdale, NJ: Erlbaum. 259–276.

Koller, D. (draft). *Virtual GIS: A Real-Time 3D Geographic Information System.* Atlanta, GA: Georgia Institute of Technology.

Lahoud, J. A., and Cleveland, D. 1994. The eyegaze eyetracking system, unique example of a multiple-use technology. In *Proceedings of the IEEE Dual-Use Technologies and Applications Conference.*

Rizzolatti, G.; Riggio, L.; Dascola, I.; and Umilta, C. 1987. Reorienting attention across the vertical and horizontal meridians: Evidence in favor of premotor theory of attention. *Neuropsychologia* 25:31–40.

Tanenhaus, M.; Spivey-Knowlton, M.; Eberhard, K.; and Sedivy, J. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268:1632–1634.

Yarbus, A. L. 1967. *Eye Movements and Vision.* New York: Plenum Press.

Young, L., and Sheena, D. 1975. Survey of eye movement recording methods. *Behavior Research Methods and Instrumentation* 7(5):397–429.