Available online at www.sciencedirect.com
SCIENCE DIRECT
Computers & Graphics 30 (2006) 619–628
COMPUTERS & GRAPHICS
www.elsevier.com/locate/cag

Technical Section

"Verba Volant Scripta Manent" a false axiom within virtual environments. A semi-automatic tool for retrieval of semantics understanding for speech-enabled VR applications

Giuseppe Conti*, Giuliana Ucelli, Raffaele De Amicis

Fondazione Graphitech, via F. Zeni 8, 38068 Rovereto, Italy

Abstract

Traditional interaction with virtual environments (VE) via widgets or menus forces users to rigidly sequential interactions. Previous research has proved that the adoption of speech recognition (SR) allows more flexible and natural forms of interaction resembling the human-to-human communication pattern. This feature though requires programmers to compile some human supplied knowledge in the form grammars. These are then used at runtime to process spoken utterances into complete commands. Further speech recognition (SR) must be hard-coded into the application.

This paper presents a completely automatic process to build a body of knowledge from the information embedded within the application source code. The programmer in fact embeds, throughout the coding process, a vast amount of semantic information. This research work exploits this semantic richness and it provides a self-configurable system, which automatically adapts its understanding of human commands according to the content and to the semantic information defined within the application's source code.

Keywords: Virtual reality; Speech recognition; User interfaces; Semantics

1. Introduction

Most virtual reality (VR) applications adopt menu-based interactions to provide access to the functionalities available within the three-dimensional (3D) environment. The interaction metaphors can vary from traditional 2D menus, to more complex 3D widgets. Hybrid approaches make use of further abstraction levels by introducing elements, such as a tablet or a pen [1–3], which in turn provide access to traditional menu-based commands. Major advances in the field of human–computer interaction (HCI) have fostered the adoption of more natural forms of interaction. In fact new interfaces have gone beyond the mere decoding of users' pointing actions by taking advantage of the information encoded through voice, gestures or gaze. This has led to multimodal VR interfaces where multiple communication channels [4] are used. Particularly, the integration of gestures and SR within VR applications can let users benefit from both the intrinsic spatial nature of VR environments and from the directness and efficacy of spoken commands. As asserted by McNeill [5] speech and gestures originate from an internal knowledge representation that encodes both semantic and visual information. Their integration becomes a decisive advantage in applications targeted to the engineering design domain. Indeed, as highlighted by Reeves et al. [6], the design experience, which is based on the generation of shapes [7], strongly benefits from the support of multi-sensorial, or multimodal, interactions. In fact, as proved by cognitive scientists, virtual reality aided design (VRAD) applications can exploit speech as a complementary conceptual channel

Abbreviations: AI, artificial intelligence; AR, augmented reality; CFG, context-free grammars; FS, feature structures; HCI, human computer interaction; NLP, natural language processing; SAPI, speech application programming interface; SR, speech recognition; TTS, text-to-speech; VE, virtual environments; WSD, word sense disambiguation; VERBOSE, voice enabled recognition based on semantic expansion; VR, virtual reality; VRAD, virtual reality aided design; XML, extensible mark-up language

*Corresponding author. Tel.: +39 0464 443450; fax: +39 0464 443470.
E-mail addresses: giuseppe.conti@graphitech.it (G. Conti), giuliana.ucelli@graphitech.it (G. Ucelli), raffaele.de.amicis@graphitech.it (R. De Amicis).

0097-8493/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.cag.2006.03.004

[8] capable of transmitting information not easily defined spatially through gestures. With respect to this Kendon [9] highlights how, during the drawing of shapes, gestures tend to co-occur with phonologically prominent words.

Concurrent exploitation of gesture and speech requires a representation of the knowledge [5] relative to the specific context domain which allows the appropriate decoding of the user's action. When such knowledge is supplied the use of speech within VR applications allows direct access to functions available within the virtual world through a more "natural" form of interaction [4] which resembles the human-to-human communication pattern [10].

However the implementation of SR capability requires the developer to manually define the necessary knowledge base. This defines the way commands detected by the SR subsystem are to be translated into actions within the VR environment. More precisely, developers typically need to manually define how spoken utterances are bound to the system's functions through an inefficient process. This process must be repeated every time a modification to the system code is introduced.

This research work tackles these issues through voice enabled recognition based on semantic expansion (VERBOSE) a system capable of automatically generating a speech-enabled interface for VR applications which intelligently adapts to the user's commands. VERBOSE plays a key factor during the development of a VR application since it can be integrated with the specific parts of the system architecture dealing with the interaction process which are hard-coded to the rest of the application.

The process proposed allows, with few minimum variations to the original application, to automatically generate both the body of knowledge required by the SR engine and to connect the SR subsystem to the VR application's functions. The system presented is capable of self-configuring to adapt its comprehension to the users' spoken commands by automatically generating the relevant context-free grammars (CFG) from the application source code. The system also generates automatically the information necessary to its text-to-speech (TTS) functionalities necessary to provide an adequate interaction level. The approach developed takes advantage of the semantics defined within the source code by exploiting the vast amount of information encoded by the developer. This information is compiled by the user when defining class, instance, field or function names or the class inheritance. The resulting combination of form-descriptive gestures used to sketch and deform models three-dimensionally, together with the adoption of a flexible speech processing functionality, delivers an improved interface which can let users explore the 3D space and access its functionalities in an intuitive and natural way.

2. Related works

Traditionally speech recognition facilities have been embedded into VR systems to provide a "natural" means of interaction [4] with the virtual world, and to enhance the efficiency in the workflow. In Ref. [11] the augmented reality (AR) system described has been developed for educational purposes, allowing control of interface components through the use of spoken commands. Most systems developed for engineering applications [12], for complex assembly and maintenance tasks [13] usually use off-the-shelf speech engines [14] to recognize short commands or in replacement of simple input from the keyboard [15]. Other works have led to the creation of speech-enabled VR/AR environments based on portable devices [16].

More advanced multimodal VR applications such as in Ref. [17] have proposed an agent-based structure capable of processing inputs from different modalities through the use of feature structures (FS) unification [18]. The unification process checks for the compatibility between data structures and it merges their features into a single data structure. In Ref. [19] the authors propose the use of interaction graphs, diagrams whose tokens contain information coming from different modality, to show the user's progress within the current task.

The StudierStube [1] VR/AR platform introduces a new level of abstraction to the multimodal interaction. The system makes use of an open architecture for tracking devices, called OpenTracker [20], which provides high-level abstraction over different tracking devices and interaction modes.

As far as spoken commands are concerned, several commercial systems [21] make use of finite state grammar, which allow filtering the number of messages to be decoded by the system. Most modern engines [14,22] make use of pre-defined dictionaries and rule sets compiled into CFGs. These are used by the system to retrieve, from spoken utterances, the information required to activate the relevant commands. Semantic information can be also included at a grammar level in various commercial recognizers [14,21]. The recognition subsystem is interfaced to the VR/AR environment and the relations between spoken commands and computer actions are hard-coded. The output of the speech recognition process is passed to a language parser that interprets it accordingly.

This approach is preferred when a relatively limited set of commands must be decoded, as in the case of VR applications. In fact this approach enhances precision [4] over standard dictation systems since it allows great narrowing of the number of commands that have to be interpreted by the system, from general natural language to a specific domain. Vocabularies are defined apriori [23], the number of recognizable commands is limited in size [24] and defined according to the specific context of the application. As noted in Ref. [23] the development of more comprehensive vocabularies and grammars represents an important achievement since this can enhance significantly the expressive power of the application.

Past works [4] stressed the need for expressing knowledge contained in VEs and for extracting semantics out of