2011 International Nuclear Atlantic Conference - INAC 2011
Belo Horizonte,MG, Brazil, October 24-28, 2011
ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN
ISBN: **978-85-99141-04-5**

TANGIBLE INTERFACES FOR NAVIGATION IN VIRTUAL ENVIRONMENTS: A ARGONAUTA REACTOR CASE STUDY

Victor Gonçalves G. Freitas[1,2,3], Antonio Carlos de A. Mol[1,2], Cláudio Márcio N. A. Pereira[1,2], Maurício da Cunha[1], Diogo Ventura Nomiya[1], André Cotelli do E. Santo[1]

[1] Instituto de Engenharia Nuclear (IEN / CNEN - RJ)
Rua Hélio de Almeida, 75
21941-906 – Rio de Janeiro, RJ

[2] Universidade Gama Filho
R. Manuel Vitorino, 553
20740-900 – Rio de Janeiro, RJ

[3]COPPE/UFRJ - Nuclear
Universidade Federal do Rio de Janeiro
Ilha do Fundão, s/n
21945-970 Rio de janeiro, RJ

ABSTRACT

This work presents a interaction system for virtual environments that allows users control platform without the need of using mouse and keyboard. Through the head movement and voice commands is possible the navigation and control of virtual human (avatar), thus, better human-computer integration. The control system called SOMI (Sounds and Motion Interface) is based on speech recognition using artificial neural networks (ANN), where once the ANN are trained for the different users it possible direct a vocal command to a command of avatar, resulting in the possibility of control by voice. Head movements are recognised using the system infrared(IR) head tracking, this system is based in a IR camera detecting the position of IR leds positioned on the user's head to put the avatar vision in accordance with the vision of the user. The SOMI was integrated into a system called Virtual Argonauta (VA), this system configures itself as a virtual platform for operations training at nuclear reactor research of the Nuclear Engineering Institute (IEN/CNEN), until the present work the virtual person was all controlled by mouse and keyboard, preventing the use of head mounted display (HMD) and decreasing user immersion. The results show an interface that promotes a more effective and engaging manner allowing the use of HMD and with that greater immersion of human beings to the virtual environment, more specifically the VA environment.

Corresponding author. Tel.: +5521 21733894.
E-mail addresses: mol@ien.gov.br (A.C.A. Mól), vgoncalves@ien.gov.br (V.G.G. Freitas), calexandre@ien.gov.br (C.A.F. Jorge), cmnap@ien.gov.br (C.M.N.A. Pereira), cotelli.andre@gmail.com (A.C.E. Santo), diogonomiya@gmail.com (D.V.Nomiya)

1. INTRODUCTION

For most of the history of computing, people have relied on screen-based text and graphics as the primary means for representing digital information. Whether the screen is desk-mounted, head-mounted, hand-held, or embedded in the physical environment, the prevailing combination of screens and general-purpose input devices has cultivated a predominantly visual paradigm of human-computer interaction.

Fitzmaurice, Ishii, and Buxton took an important step towards describing a new conceptual framework with their discussion of “graspable user interfaces” [2], Building upon this foundation, Ishii and Ullmer extended these ideas and proposed the term “tangible user interfaces” (TUIs) in them works [6]. The most popular application of tangible interfaces has been to use physical objects to model various kinds of physical systems. For example, tangible interfaces have been used to describe the layout of assembly lines [9,3], optical systems [10], buildings [11], furniture [3]. Tangible user interfaces are broadly concerned with *giving physical form to digital information*. At the highest level, there are two basic facets of this approach. First, physical objects are used as representations of digital information and computational operations. Secondly, physical manipulations of these objects are used to interactively engage with computational systems [6].

Extends the concepts of tangible interfaces this work proposes the use of voice commands and head movements to the control of virtual environments, called SOMI, the system allows a more natural integration bringing more immersion to the user. As application, proposed an extension of the work of [4] in order to control the avatar in the virtual environment research reactor called Virtual Argonauta (AV).

2. METHODOLOGY

The system developed, SOMI, uses a methodology that can be divided into two parts: i) voice recognition system. ii) head positioning system. Voice command is possible to navigate in AV, still with vocal system its possible operations system that before were performed from the keyboard, such as example: walk, pick up objects, lowering, among others. With head-tracking is possible avatar vision directional, thus having a total control over the virtual person.

2.1. Speech recognition

This system has been designed and optimized by making use of two methods: analysis of cepstrais coefficients, used on the voice signal pre-processing and neural networks, to the recognition process itself. Pre-processing of voice signals is a necessary task, because it eliminates the voids that precede and succeed the word during a recording. Once isolated the Word are extracted parameters that represent. These parameters are presented to neural networks, which then perform the recognition in one of seven commands: “abaixo”, ”acima”, ”afasta”, ”aproxima”, ”direita”, ”esquerda” e ”para”. These seven commands can be used to represent some operation on VA, not being necessary the faithful representation of the word and its operation, for example,"afasta" can start the floor of the avatar in the environment.

2.1.1. Disposal of empty

The empty spaces of a voice signal correspond to a silence, which comes at the beginning and end of recordings. Are times when the announcer is not effectively speaking, and in these hours is obtained only a background noise. To delete these snippets, the signal was segmented into ranges of 20ms. Was calculated the energy of each of these ranges, and those that produce a value less than 1% of the range of most power were eliminated.

2.1.2. Cepstral analyses

Once isolated the spoken word, is required to obtain parameters that represent this word, because these are parameters that are submitted to neural networks during the recognition stage. For understand a bit more about the process of the voice. The voice signal can be seen as the convolution between the excitation signal and impulsive response of the vocal tract (Figure 2.1.2). Once the spoken word is strongly linked to the vocal tract, the analysis on cepstral can be used, since it is a technique applied in signal deconvolution.

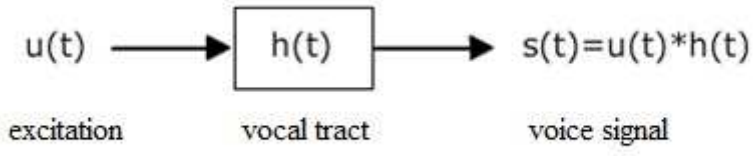

Figure 2.1.2 – simplified model of speech production

This method performs some operations on the voice signal, seeking to separate the two signals that comprise:

i-is applied to discrete Fourier transform, yielding a multiplication where before was a convolution:

$$\mathbf{s\,(w) = u\,(w)\,h\,(w)}$$

ii-take the logarithm of the module, now a sum:

$$\mathbf{\log |s(w)| = \log |u(w)| + \log |h(w)|}$$

iii-applying the inverse transform arrives to the field of frequencies (move of letters with the word frequencies), and it obtained the coefficients cepstrais:

$$\mathbf{F^{-1}\,\{\log |s(w)|\} = F^{-1}\,\{\log |u(w)|\} + F^{-1}\,\{\log |h(w)|\}}$$

$$\mathbf{cs(k) = cu(k) + ch(k)}$$

The first coefficients cepstrais are directly connected with the transfer function of the vocal tract, as the coefficients of excitation in high frequencies (with the obvious exception of the first coefficient, which also belongs to the excitation). In this way are used the 12 initial elements from the second.

2.1.3. Artificial Neural Networks

The operation of artificial neural networks (ANN) was inspired by the complex biological system of interconnected neurons. Artificial neural networks, neurons are arranged in layers intertwined. There are various types of networks, which vary according to the arrangement of layers or of links [5]. In this work were addressed networks General Regression Neural Network (GRNN). Neural networks can be seen as belonging to a topic of Artificial Intelligence, known as machine learning. This subfield is dedicated to developing methods that allow the computer to "learn" based on data made available to machinery. In the case, the neural networks have been used in pattern recognition, i.e. data parameters that characterize a word, it is hoped that the neural networks recognize that word is this.

The NN received as input the 600 cepstrais coefficients of the spoken word, and returned as output of the seven comados of voice: “abaixo”, ”acima”, ”afasta”, ”aproxima”, ”direita”, ”esquerda” or ”para”. Before the machine can recognize effectively the words, it must go through a learning process. Here were used the supervised learning method, which can be understood as if there was the presence of a teacher that taught the machine each of the seven target words. In this method, to promote the learning of each standard, are presented to machine the coefficients cepstrais of various circumlocutions and indicated which voice command they represent. These data are called the training set, because it is from them that neural networks are trained.

The networks were trained with 15 phrases each command (some with 30), generating a training set with 105 recordings (or 210). Already at the stage of recognition, after receiving the 600 cepstrais coefficients of a Word, the nets return seven values from 0 to 1, representing the seven data sets, corresponding to seven commands. The corresponding command to the largest value returned by the networks is the command effectively recognized from the word said. Ideally, only one of the values must be close to 1, while the other six must be close to 0.

With this methodology was possible recognition of the voice of a user, just so, just the prior training of neural networks with the vocal tract of each user. Once the system is already trained and recognizing the voice of the user, these commands are exchanged for keys which command the avatar in Virtual Argonauta (VA).

2.2. Head-Tracking

For users who cannot control a computer through a standard keyboard and mouse, alternative interfaces have been developed. There are a number of commercially available devices that act as ‘Head- Tracking Mouse’, this system aims to control the mouse using the position of the head [1]. Some are specifically designed for people with disabilities, others have recently been developed for the games market. The first answer to the question is the cost of those systems, specifically aimed for people with disabilities, they can cost from US$ 100 until US$ 1000. The system used in this work cost US$ 150. What if demand in this method is

the information of six degrees of freedom, 6DOF (Figure 2.2.1), the heads up on camera. Three are relative head rotation in three axes (Euler angles) and three other relative displacement to the axes.

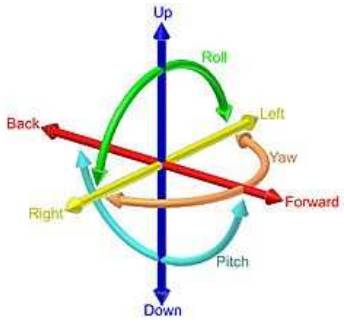

Figure 2.2.1 - 6DOF

Two parameters were used for the movement of the head: rotating on Y axis (yaw), rotation on the X axis (pitch). This method uses an infrared camera (TRACKIR5) owner, sold by the company Natural Point, with a cost of $ 150. This camera has infrared leds, which serve to illuminate the environment and a device with three points was placed on the helmet in such a way as to reflect this wave that comes to be triangularized, so we detect the rotation of the head. The reflection of waves IR utilizes the same methodological principles and the same code of the system where the leds are placed on the head, but the reflection of the wave is more portable than the use of leds infrareds in helmet, both were tested for the positioning.

3. APPLICATION

The application of the methodology described in this work was held at Virtual Argonauta (Figure 3.1) that have been developed over the works of [7] and [4]. This works UnrealEngine2 Runtime Demo Version (EpicGames) has been modified and adapted to perform virtual simulation of environments, with this application in mind, and results are shown. This game engine has its own functions already implemented, related to navigation in virtual environments, considering physical laws of movement, and collision with scenario and objects. Thus, these functions can be used for virtual simulations of any other scenario and it was aplicated on research reactor at IEN/CNEN called Virtual Argonauta (Figure 3.2). Further, UnrealEngine2 Runtime Demo Version has been chosen because it is a demo version that can be used for non-commercial and educational purposes, with source code available, that can be adapted to create new scenarios, and to include new functionalities, depending on the application's needs. This work has been developed at Laboratório de Realidade Virtual – LABRV [8], Instituto de Engenharia Nuclear – IEN/CNEN.

Figure 3.1 - Virtual Argonauta

Figure 3.2 - Argonauta research reactor - IEN/CNEN

4. RESULTS

As the results of applying the methodology its possible to control the avatar from the voice command of the user and the avatar vision targeting according to moving user head with HMD.

4.1 - Head-tracking results

Table 4.1 shows the results refentes the location of the head, were taken 100 samples classified in 10 sets of experiments, the position the Centre refers to head stop focused on the Centre, the left edge refers to head stop focusing on the left of the monitor, the right edge refers to head stop focusing on the right edge of the monitor and the average of each set means the average of 100 output values of the position with the head stop in each point. The standard deviation is the variation of measurement even with the head stop and a specific focus.

Table 4.1- Results from the variation of output of the positioning of the head.

	Centre			Left			Right	
	Average	Standard Deviation		Average	Standard Deviation		Average	Standard Deviation
1	0,508	0,200		-26,720	0,190		24,369	0,097
2	1,306	0,177		-22,152	0,134		27,679	0,139
3	0,256	0,237		-20,221	0,151		27,696	0,125
4	-0,758	0,214		-25,202	0,156		28,001	0,129
5	0,566	0,291		-21,482	0,078		26,597	0,151
6	-1,179	0,263		-17,064	0,115		26,966	0,172
7	-0,153	0,183		-18,288	0,135		28,669	0.127
8	0,015	0,155		-19,026	0.087		24,890	0,159
9	-0,744	0,140		-22,315	0,083		26,795	0,115
10	0,657	0,211		-22,036	0,170		20,284	0,180

4.2 - Voice recognised results

During testing, a set previously known phrases is sent as input to the system. This, from templates, performs recognition as if the phrases were unknown. This way, it is returned to the recognized string of words, which is then compared with the sequence of words represented by the circumlocutions. In this context, one of the most used metrics to condone the recognition systems performance is the rate of errors in recognition (WER-word error rate), which is defined as:

$$\mathbf{WER = (S + I + D).N^{-1}}$$

Where S, I, and D are respectively the number of errors by words replaced (S), insert (I) and deleted (D), where N is the total number of words to be recognized. The other metric commonly used is the word recognition rate (WRR-word recognition rate) that, unlike the previous, bigger is better. This work uses the WRR, whose definition is given by:

$$\mathbf{WRR = 1 - WER}$$

Another important measure is the real time factor (RTF – real time factor) or real-time speed. The RTF is a way to evaluate how fast the system is able to perform the task of recognition, being highly dependent on the hardware used. The RTF is defined by:

$$\mathbf{RTF = TP.TA^{-1}}$$

Where TP is the equivalent of the audio processing time and TA is the time (duration) of input audio. This way, real-time systems must have an RTF as low as possible. For systems with limited vocabulary, is surprisingly a RTF greatly reduced. To get a more accurate

estimate of system performance, these metrics were used in conjunction with the technique of cross-validation. In it, the data set is divided into k equal lengths segments, being then made k experiments. Each time, only one of the segments is used to test, while the other k-1 are used for training. Cross-validation is illustrated by Figure 4.2.1. In this work, were made four experiments (k = 4), and measure the WRR of each one of them. This way, the WRR estimated amounts to the average of the WRR of each experiment.

$$\mathbf{WRR} = \Sigma^{k}_{i=1} \mathbf{WRR}_i \,.\, k^{-1}$$

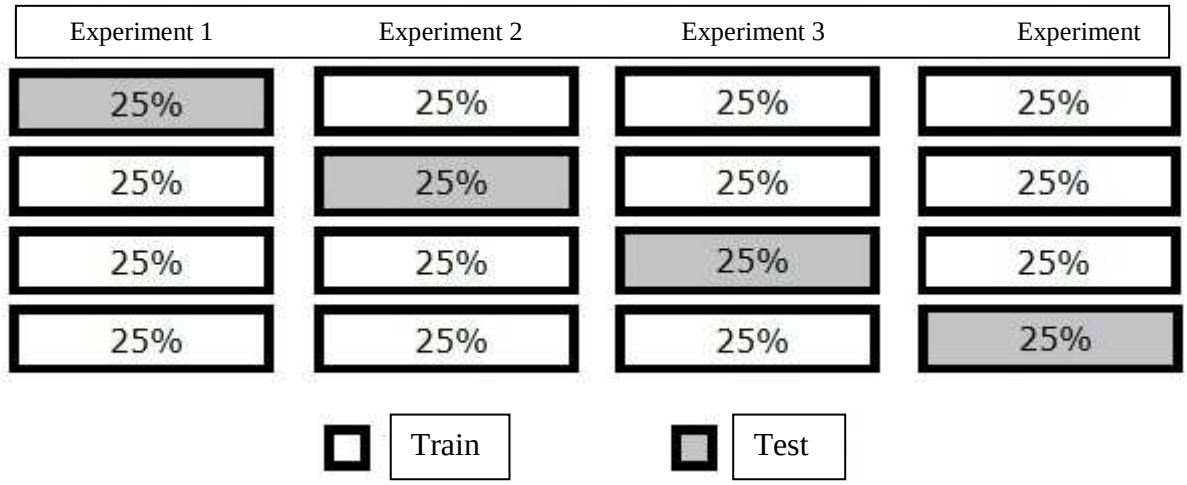

Figure 4.2.1- Cross-Validation

The tests were run on a computer with Intel architecture CoreTM2 Duo 2.80 GHz with 2 GB of RAM. Results can be seen in table 4.2.1 and 4.2.2.

WRR_1	WRR_2	WRR_3	WRR_4
94.5%	98.2%	97.1%	93.5%

Table 4.2.1- Results from cross-validation

RTF_1	RTF_2	RTF_3	RTF_4
0.026	0.023	0.023	0.025

Table 4.2.2 - Results from cross-validation

Thus, their estimated values equal to the average of the results obtained in each experiment:

WRR = 95:8%
RTF = 0:024

From the voice command is possible around the avatar by virtual environment using the seven commands trained neural networks as described in section 2.1.

4.3 - Application Results

Figures 4.3.1 shows the HMD with three reflectors that make head-tracking system work and Figure 4.3.2 shows the complete system, SOMI, mounted next to the user, for control of the VA. The image seen on the monitor of Figure 4.3.2 is the same image viewed on HMD.

4.3.1 - HMD with IR reflectors.

4.3.2 - SOMI operating VA.

5. CONCLUSIONS

The extension of the possibilities of integration between users and computer systems increasingly decreases the Execution Gulf, the Gulf which is understood as the gap between the goal of the user and the execution of the tasks in the system, thus, the efficiency in the control of operating platforms as Virtual Argonauta. Through these new devices and techniques for integrating the environment can now be operated without the use of mouse and keyboard. With these new integration techniques, called tangible interfaces, it is possible also benefit users with motor impairments, in a manner not detrimental to the training of its operations in the virtual environment of nuclear reactor research Argonauta.

ACKNOWLEDGMENTS

This research was sponsored by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro - FAPERJ and Conselho Nacional de desenvolvimento Científico e Tecnológico - CNPq. Our thanks also to the Comissão Nacional de Energia Nuclear - CNEN, that, through the Instituto de Engenharia Nuclear - IEN, has provided all necessary resources to the development of this work.

REFERENCES

1. Evans, D. G.; Pettit S.; and Blenkhorn P., "A head operated 'Joystick'," in Proc. 5th Int. Conf. Comput. Helping People Special Needs, R. Oldenbourg, Ed., 1996, pp. 85–91.
2. FITZMAURICE, G., ISHII, H., AND BUXTON, W. (1995). Bricks: Laying the Foundations for Graspable User Interfaces. In *Proceedings of CHI'95*, pp. 442-449.
3. FJELD, M., BICHSEL, M., AND RAUTERBERG, M. (1998). BUILD-IT: An Intuitive Design Tool Based on Direct Object Manipulation. In *Gesture and Sign Language in Human-Computer Interaction, Lecture Notes in Artificial Intelligence,* v.1371, Wachsmut and Fröhlich, eds. Berlin: Springer- Verlag, pp. 297-308.
4. FREITAS, V. G. G. ; PEREIRA, C. M. N. A. ; MÓL. A. C. A. ; ALEXANDRE, C. . Radiation dose rate map interpolation in nuclear plants using neural networks and virtual reality techniques. Annals of Nuclear Energy JCR, 2010.
5. Haykin S. Neural Networks and Learning Machines. Third Edition - New York : Prentice Hall, 2009. - 936p. ISBN/ISSN 978-0-13-147139-9
6. Ishii, H. (2008). The tangible user interface and its evolution. Communications of the ACM, Volume 51 Issue 6.
7. MOL, A ; JORGE, C ; COUTO, P ; AUGUSTO, S ; CUNHA, G ; LANDAU, L . Virtual environments simulation for dose assessment in nuclear plants. Progress in Nuclear Energy JCR, v. 51, p. 382-387, 2009.
8. Mól, A.C.A., Grecco, C.H.S., Carvalho, P.V.R., Oliveira, M.V., Santos, I.J.A.L., Augusto, S. C., Viana Filho, A.M., August 28–September 02, 2005. Implementation of the immersive virtual reality laboratory in Nuclear Engineering Institute. In: 2005 International Nuclear Atlantic Conference – INAC 2005, Santos.
9. SCHÄFER, K., BRAUER, V., AND BRUNS, W. (1997). A new approach to human-computer interaction synchronous modelling in real and virtual spaces. In *Proceedings of DIS'97*, pp.335-344
10. UNDERKOFFLER, J., AND ISHII, H. (1998). Illuminating Light: An Optical Design Tool with a Luminous-Tangible Interface. In *Proceedings of CHI'98*, pp. 542-549.
11. UNDERKOFFLER, J., AND ISHII, H. (1999). Urp: A Luminous-Tangible Workbench for Urban Planning and Design. In *Proceedings of CHI'99*, pp. 386-393.