Gesture Variants and Cognitive Constraints for Interactive Virtual Reality Training Systems

Stephanie Huette[1], Yazhou Huang[2], Marcelo Kallmann[2], Teenie Matlock[1], Justin L. Matthews[1]

[1]School of Social Science, Humanities, and Arts
[2]School of Engineering, University of California
Merced, California 95343
{shuette, yhuang6, mkallmann, tmatlock, jmatthews}@ucmerced.edu

ABSTRACT
Two studies investigated the nature of environmental context on various parameters of pointing. The results revealed the need for extreme temporal precision and the need for efficient algorithms to parse out different styles of pointing. Most variability in pointing came from individual differences, and a method to classify the kind of point and derive its temporal parameters is discussed. These results and methods improve the pragmatism of virtual reality, making events appear more realistic by emphasizing temporal precision.

Author Keywords
Virtual agents, pointing, gesture modeling, virtual reality

ACM Classification Keywords
Gesture, virtual characters, virtual reality, interactive user interfaces, training systems

General Terms
Algorithms, Experimentation.

INTRODUCTION
The system being developed is an immersive virtual reality training program. It is based on interactive virtual human agents [2] and makes use of realistic, parameterized gestures. To understand the constraints of how a virtual agent is perceived naturally, it is becoming increasingly important to be able to generalize types of gestures as a function of context. This requires extensive real human-subjects testing of when, where, and how certain parameters are used in everyday gestures. For example, people may point very quickly and precisely at a pair of scissors in front of them, but point slowly and vaguely at a building located 50 meters away. Our system contains two phases that target virtual training. In the modeling phase, experts in the training subject could model the needed gestures and actions in a straightforward way by directly demonstrating them with motion capture devices and without the need of having previous experience with the system. In the training phase, the captured example motions are then re-used by the virtual human to train apprentice users. In particular, reproduced motions are parameterized with respect to arbitrary target locations in the environment. Figure 1 presents one typical scenario modeled by our system, where a parameterized computational gesture model is needed in order to allow the virtual human to point to any desired feature on equipment and explain its function. Our training set-up allows users to immersively interact with virtual characters in a variety of training scenarios. A full-scale immersive experience with virtual humans has the potential to improve learning in many ways, analogously to how people learn from each other. Figure 2 shows the modeling phase and training phase for pouring water actions.

Figure 1. Example of a training application where a parameterized computational gesture model is needed in order to allow the virtual human to point to any desired feature in a given equipment and explain its function.

In the human behavioral domain, gesture research is typically reliant on hand coding of videotaped observations, and various coding schemas used to parameterize observed behaviors [4]. These methods have yielded an immense wealth of knowledge about how gestures are evoked and what their underlying cognitive underpinnings are (e.g., [5]).

IUI 2011, February 13–16, 2011, Palo Alto, California, USA.

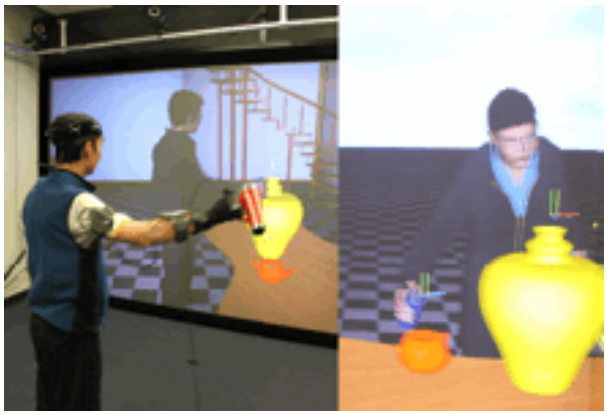

Figure 2. Our training set-up allows users to immersively interact with virtual characters in a variety of training scenarios. Full-scale immersive experiences with virtual humans has the potential to improve learning in many ways, in analogy to how people learn from each other.

However, these methods have not enabled generalizable parameters that adequately reflect a reasonable range of human movements. Note that humans are highly attuned to subtle differences in motor behavior, including minor shifts in velocity in cursive handwriting [6]) and pointing velocity related to the size of an object, as in Fitts' law [7]. In brief, it is challenging to fit scalable parameters onto virtual agents that we perceive as moving naturally, especially in light of the time-varying dynamics of motor movement (e.g., velocity profiles as opposed to constant velocity), the degrees of freedom in the upper limbs, and the meaningful variability such as 1/f noise found in distance between footsteps in healthy human subjects [3].

The execution of motor movements by humans is a clear, predictable science, and many advances in the creation of laws governing these movements have been made in recent years. The 2/3 power law governs elliptical motions made on a plane, e.g. drawing an elliptical shape on a plane [9]. This model effectively establishes a relationship between angular velocity and curvature of the end-effector trajectory. The critical feature of this law is that velocity is not constant, and that humans can detect incrementing or decrementing velocities at specific points on an ellipse. If constant velocity is used, the motion will then appear artificial and unnatural. There are also laws governing choice of trajectory, where the motor system will choose the trajectory of least resistance. In terms of parameters, this is described by choosing the smoothest path, thus, minimizing jerk (the derivative of acceleration) [8].

Equally important, if not more important than systematic variability and correct parameterization, is the context in which they occur. In terms of these aforementioned laws, Fitt's law is only followed by small pointing gestures, and the 2/3 power law is only followed with planar motions, e.g. handwriting. In recent years, contextual manipulations have been shown to strongly affect behavior. Humans can easily detect artificiality when parameters are not an optimal fit with a given context. Thus, the job of modeling virtual agents is complex and challenging, especially because a given parameterization must find environmental contexts in which a certain behavior can be generalized. An example of this can be seen in the current research. For example, people will hold a point longer when they are pointing at an object farther away, where the proposed explanation is understood in a communicative context: it takes longer to follow a point trajectory when the target object is further away [11].

Researchers interested in gesture have investigated how people point in various contexts to address some of these concerns [12]. In addition to these experiments, one of the most important factors to achieving natural looking gestures was determined to be related to a coherent velocity control. This paper discusses the initial stage of our work on gesture and virtual characters. The aim is to build a computational model for gesture parameterization with applicability to virtual environments.

Research on human participants

The purpose of the following studies was to explore the domain of everyday gestures, and ultimately, to advance theory and practice in the gesture research. Towards this end, we integrated contextual factors and observed which salient parameterization differences appeared to be key to making the gesture as natural as possible.

Study I

In our first task, 25 participants were videotaped while they performed a series of simple pointing tasks. These pointing tasks assessed consistency (and variability) across pointing behaviors and individuals to help set the stage for motion capture research. These early studies, intended to help establish parameters of pointing, analyzed factors including velocity of the hand, acceleration, distance from object, and hold pointing. Also considered was whether the person utilized singular versus plural numbers of objects. This manipulation instructed people to either "Where is the stapler?" or "Where is the stapler and the pencil?" in the plural condition. Position of the body was also manipulated between subjects. Specifically, half of the participants remained seated and the other half remained standing. All participants were University of California, Merced undergraduates enrolled in cognitive science or psychology courses who volunteered for extra credit. All signed provided consent forms to be videotaped and adhered to IRB approval.

Design

In all studies, participants sat or stood at a table that contained basic objects, primarily office supplies, such as staplers, paper clips, and scissors. The first study required participants to point at objects arranged in an array. Participants were asked in the format "Where is the [x]?" for each object in the array, to investigate if position

influences the kind of point made. Next, participants were asked to point consecutively to two objects, as in "Where is the [x] and the [y]?" where similarity of the object (e.g. small pair of scissors and a large pair of scissors) was also manipulated. The size manipulation controlled for distance by having objects either close or far away from one another, and objects either similar or dissimilar.

This design yielded a 2x2 within subject manipulation for the trials during which they pointed to two objects (size x similarity), and a between subjects manipulation of body position.

Figure 3. Pointing gesture performed by human participant

Trends observed

In general, there was consistency in pointing at various types of objects (e.g., a stapler versus a paper clip). And overall, each participant appeared to use a relatively steady velocity, similar distance to object, and length of hold for each trial, regardless of the target object. However, there were noticeable differences in the length of time a pointing gesture was held. The more distant the pointing, the longer it was held ($p=.003$). There were also differences in distance (close vs. far from object). About 30% of the time, participants touched the object; 66%, they were close; and 4%, they were far. And when people pointed at two objects in a row, the second point tended to be faster and held for a shorter duration.

Study II

A follow up study was conducted in which participants performed iconic and pointing gestures. They were video taped while they performed their prototype of a controlled, precise point, as well as gestures relating to health care instruction (e.g., attending to an injury). Example prototypes included simple actions such as cleaning a wound, opening a bandage, and applying pressure. These videos were analyzed in order to find prototypical gesture variables and to pilot this explicit mode of instruction. Participants were issued instructions such as "Please show how you would clean a wound with one hand, using a circular motion."

Even though participants were explicitly instructed to perform actions, substantial variability was observed across individuals, revealing the need for clear methods to parse out different categories with limited variability, to make virtual agents look realistic. This framework adjusts data collected from multiple people, and post-hoc performs a cluster analysis. This statistical analysis is used in many areas such as machine learning, image analysis, and in pattern recognition. For example, after motion capture, an algorithm can be applied which analyzes the velocity profile, and would output that the person is a fast pointer. Other motoric characteristics of fast pointers can then be applied, where the fast-pointer cluster is near a cluster of minimal jerk, short end-point hold time. This avoids averaging the fast pointer with the slow pointer, affording much more realistic, and individualized pointing from the virtual character, which will give various avatars the appearance of having a more specific personality.

Discussion & Directions

At this stage, we are interested in whether observed variability is related to differences in cognitive processing, or whether it is simply a matter of noise that can be neglected in modeling pointing and gesture behaviors. In other work, we are addressing how this variability can be manipulated by different words that evoke semantic differences during language comprehension. This will provide insights into which kinds of parameters are negligible and can be simply be treated as noise, and which are specifically related to systematic differences in pointing.

The goal of future studies is thus to discover underlying representational differences that directly map onto abstract features of the environment. It will also be possible to derive equations that explicitly define the velocity profile of a point to an object thought of in terms of a whole, as well as one describing objects thought of in terms of their components. For instance, when asked to "point to a telephone" this may be different from "point to the number 5 on the phone". In other words, pointing to a feature of an object may be much different than pointing to the whole object itself. This is also a property of the language, and thus may be generalizable to componential / holistic environmental properties in addition to the way the linguistic structure modulates the behavior produced.

More generally, we observed that for a virtual reality training system to be believable, it must be precise to a millisecond timescale. In one study on haptic-audio synchrony, participants were asked to strike a brick with a rubber tipped hammer, and a computer generated either a synchronous, or asynchronous percussive sound [1]. Participants were able to detect the asynchrony with a 24ms sound onset latency. In the visual system alone, people are able to detect asynchrony in visual stimulus onsets that occur only 3-4ms apart [10]. This degree of sensitivity must be taken into account in programming, as hardware delays and inefficient algorithms will produce unrealistic looking animations.

CONCLUSION

The present research focused on observing and categorizing kinds of pointing based on environmental parameters. The main results included a relationship between end-point hold

time and distance to the target, and finding that individual subjects are consistent, but when pooled exhibit too much variability to be averaged together and look realistic. The job of creating a virtual reality is not simply re-instantiating physical parameters on the virtual level. The human perception of reality is not so much about where something is or what it looks like, as when and how it happens.

This information is being incorporated into a virtual-reality training system (Figures 2 and 3) that aims to be precise on the millisecond time-scale, in order to achieve training systems based on virtual humans as realistic and effective as possible.

ACKNOWLEDGMENTS

We would like to thank research assistants David Sparks and Jeremy Hunter for help with data collection and coding. This work was supported by NSF awards CNS-0723281 and IIS-0915665

REFERENCES

1. Adelstein, B., Begault, D., Anderson, M., & Wenzel, E. Sensitivity to haptic-audio asynchrony. *International Conference on Multimodal Interfaces*, (2003), 73-76.
2. Camporesi, C., Huang, Y., & Kallmann, M. Interactive motion modeling and parameterization by direct demonstration. *Intelligent Virtual Agents (IVA), 2010.*
3. Hausdorff, J., Balash, J., & Giladi, N. Effects of cognitive challenge on gait variability in patients with parkinson's disease. *Journal of Geriatric Psychiatry and Neurology, 16* (2003), 53-58.
4. Kendon, A. Gesture: Visible Action as Utterance. Cambridge, UK: Cambridge University Press (2004).
5. McNeill, D. Gesture and Thought. Chicago, IL: The University of Chicago Press (2005).
6. Orliaguat, J., Kandel, S., & Boë, L. Visual perception of motor anticipation in cursive handwriting: Influence of spatial and movement information on the prediction of forthcoming letters. *Perception, 26* (1997), 913-928.
7. Soukoreff, W. & MacKenzie, S. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *International Journal of Human-Computer Studies, 61* (2004), 751-789.
8. Schaal, S. & Sternad, D. Origins and violations of the 2/3 power law in rhythmic three-dimensional arm movements. *Experimental Brain Research, 136* (2001), 60-72.
9. Viviani, P. & Terzuolo, C. Trajectory determines movement dynamics. *Neuroscience, 7* (1982), 431-437.
10. Westheimer, G. & McKee, S. Spatial configurations for visual hyoperacuity. *Vision Research, 17* (1977), 941-947.
11. Bangerter, A., & Oppenheimer, D.M. Accuracy of detecting referents of pointing gestures unaccompanied by language. *Gesture, 6* (2006), 85–102.
12. Kipp, M., Neff, M. & Albrecht, I. An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation, 41* (2007), 325-339.