Methodologies for the User Evaluation of the Motion of Virtual Humans

Sander E.M. Jansen[1,2] and Herwin van Welbergen[3]

[1] Department of Computer Science, Utrecht University, The Netherlands
[2] TNO Human Factors, The Netherlands
[3] Human Media Interaction, University of Twente Enschede, The Netherlands

Abstract. Virtual humans are employed in many interactive applications, including (serious) games. Their motion should be natural and allow interaction with its surroundings and other (virtual) humans in real time. Physical controllers offer physical realism and (physical) interaction with the environment. Because they typically act on a selected set of joints, it is hard to evaluate their naturalness in isolation. We propose to augment the motion steered by such a controller with motion capture, using a mixed paradigm animation that creates coherent full body motion. A user evaluation of this resulting motion assesses the naturalness of the controller. Methods from Signal Detection Theory provide us with evaluation metrics that can be compared among different test setups, observers and motions. We demonstrate our approach by evaluating the naturalness of a balance controller. We compare different test paradigms, assessing their efficiency and sensitivity.

Keywords: Evaluation of Virtual Agents, Naturalness of Animation.

1 Introduction

Virtual humans (VHs) are employed in many interactive applications, including (serious) games. The motion of these VHs should look realistic. We use the term *naturalness* for such observed realism [1]. Furthermore, VH animation techniques should be flexible, to allow interaction with its surroundings and other (virtual) humans in real time. Such flexibility is offered by procedural animation methods (for example [2,3]) and animation steered by physical controllers (for example [4]). We are interested in finding methods to evaluate the naturalness of motion generated by these techniques.

Both procedural models and physical controllers typically steer only a selected set of joints. Whole body movement is generated by a combination of different procedural models and/or physical controllers that run at the same time. Such whole body involvement is crucial for the naturalness of motion [1]. It is hard to determine exactly what motion model contributed to the (un)naturalness in motion generated by such a mix of controllers and/or procedural models. We propose the use of a mixed motion paradigm [5] to augment the motion generated by a single controller on a selected set of joints with recorded (and thus assumed natural) motion on the remaining joints, in a physically coherent manner.

Zs. Ruttkay et al. (Eds.): IVA 2009, LNAI 5773, pp. 125–131, 2009.

1.1 Motion Used to Demonstrate Our Approach

We demonstrate our approach by testing the naturalness of a balance controller (based on based on [4]) that acts on the lower body. This controller balances the body by applying torques to the ankles, knees and hips. We augment this motion with a motion captured recording of an actor clapping his hands at different tempos. These recordings are applied to the arms, neck and head. A mixed motion paradigm method [5] is used to couple the two motions: we calculate the torques generated by the arms and head from the motion capture specification, using inverse dynamics. These torques are then applied to the trunk, whose movement is physically simulated by the balancing controller. To asses the naturalness of the balance controller, we compare the following motor schemes:

- Motion 1: full body mocap of the original clapping motion
- Motion 2: upper body mocap + lower body balance model
- Motion 3: upper body mocap + no movement on lower body

Our evaluation is intended to answer questions like:

- Is Motion 1 rated as more natural than Motion 2 and Motion 3?
- Is Motion 2 rated as more natural and harder to discriminate from Motion 1 than Motion 3?

1.2 Selecting a Test Paradigm

Ideally, a test-paradigm would be efficient (needing only small number of participants to get significant results) and scalable, that is, provide metrics that can be compared with metrics obtained in previous tests. The measure d' from Signal Detection Theory (SDT) [6] is used in all our tests as a scalable measure of discriminability. We define the sensitivity of a test (given certain test conditions) as d'. An efficient test-paradigm has a d' with a low variance in each test condition and large differences between d'-s measured in different test conditions. We compare the d' and variability of d' for the following test-paradigms (see section 2 for a detailed description of procedures and analysis for each of these methods):

- 2 Alternative Forced Choice (2AFC): In each test item, participants viewed two short movie clips of animated characters in succession. Each time, one of the clips was driven by motion 1 and one by either motion 2 or motion 3. The task was to decide which of these showed natural human motion.
- Yes/No: Participants viewed one clip per test-item. The movement of the VH was controlled by either motion 1 or motion 2. They were asked to indicate if the movement was based on real human data or a computational model.
- Rating: Participants viewed one clip per test item. Movement was controlled by either motion 1, motion 2 or motion 3. They were asked to rate the naturalness of the movement on a scale of 1-10 (not at all - very much).

2AFC is commonly used to evaluate the quality of animation, using a direct comparison to a reference motion [7,8]. 2AFC discourages bias and does not suffer from contextual effects [6]. However, for some animations providing such a reference motion is impractical (extra motion capture recordings are needed, the mocap actor might no longer be available, it might be difficult to record motion with the exact conditions used in the model, there might be large stylistic differences between the reference and the modeled motion, etc). For the evaluation of such animations, Yes/No is an interesting testing alternative. Using a rating method [9] allows for a direct measurement of naturalness, rather than the indirect assessment (human/model) provided by the other methods.

Question 1. Do Yes/No and rating have a higher $var(d')$ and a lower variability between the d'-s measured in different test conditions than 2AFC?

We expect 2AFC to be more sensitive than Yes/No and Rating because each test provides a direct reference to compare to. Macmillan and Creelman [6] propose using a $\frac{1}{\sqrt{2}}$ correction factor for d' obtained by a 2AFC test so that its value can be compared with the value of a d' obtained by a Yes/No test. They note that different values of this correction factor are found empirically.

Question 2. Is there a relationship between the sensitivities of the different test-paradigms?

2 Methods

2.1 Participants

29 participants (25 male, all between 24 and 52 years of age) took part in this experiment. All were free from any known neurological disorders as verified by self-report. Experience with motion capture and creating animations varied from 'not at all' to 'very much', creating a diverse group of participants.

2.2 Stimuli

18 separate clips were used during the experiment. One for each combination of the variables *motion input* (3), *viewing orientation* (2, see Fig. 1) and *clapping frequency* (3). Each clip showed a scene of a virtual human clapping its hands with a speed of 50, 110, 180 claps/minute. Motion was controlled as described in section 1.1 The runtime of each clip was approximately 4 seconds. Stimulus presentation and data collection was performed with Medialab v2008. [1].

Embodiment. We project the motion captured human movement onto the same embodiment as the VH. The fingers and face do not move in our experiment. To make sure that the unnaturalness of these unmoved body-parts does not dominate the naturalness judgment, we have selected an embodiment for our

[1] `http://www.empirisoft.com/medialab.aspx`

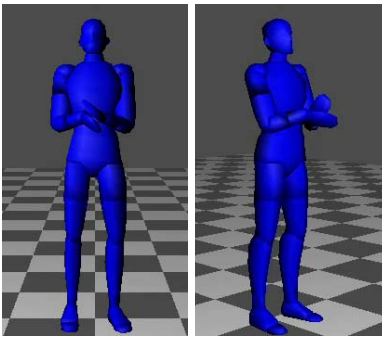

Fig. 1. Frontal and off-axis view used in the experiment

VH with simplified hands and minimal facial features (see Fig. 1). The physical model of our VH consists of 15 rigid bodies, connected by 14 joints. Each of the rigid bodies is represented by a mesh shaped as the corresponding body part in our motion captured actor. We determine the mass, CoM and inertia tensor of these bodies in a similar manner as in [5]. The physical body of the VH has roughly (within 5kg) the same total mass as our motion captured actor.

2.3 Design and Procedures

The experiment consisted of three sessions. The 2AFC and Yes-No sessions concern the discrimination between different motion inputs and the rating session required the users to rate the naturalness of a given animation. To counter learning effects, the order of the two discrimination sessions was randomized with the rating session always in between them. This was done because participants needed a certain amount of practice before they could come to a reliable rating. At the beginning of each of the three sessions, instructions were given and two test items were provided to familiarize participants with the procedure. At the end of each session, they were asked to describe what criteria they used to make their decision. A short description of each of the sessions is given in section 1.2.

2.4 Statistical Analyses

To analyze the rating data (naturalness score of 1-10), we performed a 3 (motion input) x 2 (viewing orientation) x 3 (clapping frequency) full factorial analysis of variance (ANOVA).

Signal Detection Theory is used to determine the sensitivity and variance of sensitivity for each of the *test paradigms*, (Yes/No , 2AFC and rating), *viewing orientations* (off-axis vs frontal) and *clapping frequencies* (50, 110 and 180 bpm). *d'* is a measure of perceptual difference between two observations that is not influenced by response bias (that is, a general tendency by subjects to favor the selection of one class over another) and that is comparable between different tests [6]. $d' = 0$ indicates that two observations cannot be discriminated, $d' = 4.65$

is considered an effective ceiling that indicates near perfect discrimination. d' is given by

$$d' = z(H) - z(F) \tag{1}$$

where H is the hit rate, F is the false alarm rate and z is the inverse of the normal distribution function. In the Yes/No paradigm, $H = P(\text{“human”}|human)$ and $F = P(\text{“human”}|model)$. In the 2AFC test $H = P(\text{“human left”}|humanleft)$ and $F = P(\text{“human left”}|humanright)$. Note that we do not employ the $\frac{1}{\sqrt{2}}$ correction factor for d' in 2AFC, since we are interested in determining whether the relation between d'-s found by 2AFC and Yes/No in similar test conditions is captured by this factor or any other linear relationship. The variance of d' is given by

$$var(d') = \frac{H(1-H)}{N_2(\phi(H))^2} + \frac{F(1-F)}{N_1(\phi(F))^2} \tag{2}$$

With N_2 the number of mocap trails, N_1 number of 'model' trails and $\phi(p)$ the height of the normal density function at $z(p)$.

For the rating test, we choose the area under the receiver operating characteristic (ROC) curve A_z as a measure for sensitivity (see [6], chapter 3). A_z and its variance are calculated using ROCKIT. [2]

3 Results

3.1 Comparing Motion Inputs

Motion input has a significant effect on naturalness ratings $F(2, 56) = 18.357, p < 0.001$. Tukey post-hoc analysis shows that motion input 1 was rated as more natural than motion input 2 ($p < 0.001$) and motion input 3 ($p < 0.001$). The average rating for motion input 2 although higher, was not significantly different from that of motion input 3 ($p = 0.12$).

Participants can discriminate between motion 1 and motion 2 as well as between motion 1 and motion 3 for all tests and conditions ($d' \neq 0$, $p < 0.05$). Subjects incorrectly identified motion 2 as human only in the yes/no test at 50bpm, off-axis. All other d' scores show that the subjects correctly identify motion 1 as human and motion 2 and motion 3 as nonhuman.

In the 2AFC test, subjects can discriminate between motion 1 and motion 3 significantly better than between motion 1 and motion 2 at all off-axis views and the 50bpm front view ($p < 0.001$). The rating test shows only significantly better discrimination of motion 1 and motion 3 versus motion 1 and motion 2 for the off-axis view at 50bpm ($p < 0.05$). These and other significant rating and discrimination results are illustrated in Fig. 2.

3.2 Comparing Evaluation Methods

No significant differences in $var(d')$ are found between the test paradigms. The variance between d'-s in the different test conditions is 0.95 for 2AFC, 0.25

[2] `http://xray.bsd.uchicago.edu/krl/KRL_ROC/software_index6.htm`

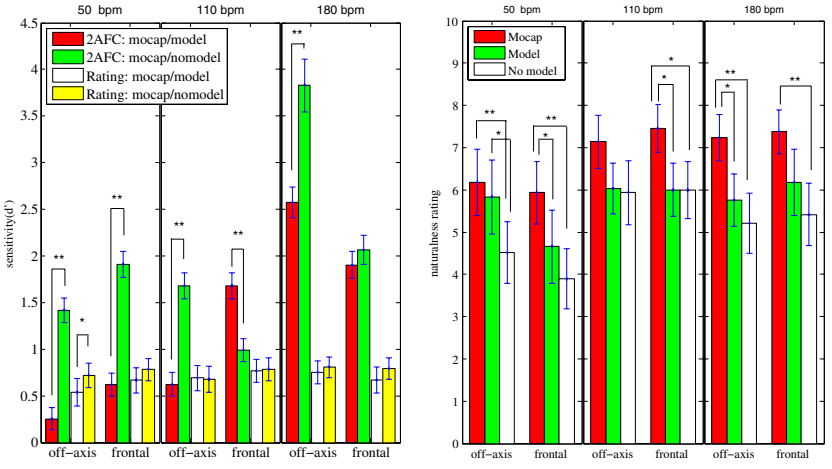

Fig. 2. Left: sensitivity as function of *viewing orientation* and *clapping frequency* for the 2AFC and rating tests. Right: mean naturalness ratings as a function of *motion input*, *viewing orientation* and *clapping frequency*. Vertical bars denote the 95% confidence intervals and. Significant differences are illustrated by * ($p < 0.05$) and ** ($p < 0.001$).

for Yes/No and 0.0058 for the rating paradigm. These values are significantly different ($p < 0.001$). We conclude that, for our test conditions, 2AFC is the most efficient test and that rating is not a good test for discrimination.

There is a strong correlation between the d' values obtained by the rating test and the d' values obtained by the Yes/No test (Pearson's $\rho = 0.906$, $p < 0.05$). Possibly the rating test as it was used in our experiment, was experienced by the subjects as a Yes/No test with an expanded grading scale. The correlation between 2AFC and Yes/No was strong ($\rho = 0.785$), but only marginally significant ($p = 0.064$). The correlation between 2AFC and rating was moderate ($\rho = 0.665$, $p < 0.05$). Significant observations made with the different test paradigms generally agreed, with the exception of the one specific case mentioned above.

4 Discussion

We have demonstrated the applicability of a mixed paradigm animation technique to evaluate physical controllers in isolation. Setting up such an evaluation is relatively easy, because the used mixed paradigm technique integrates with any existing physical simulation environment used to animate VHs [5].

Differences in variability between the d' for different test conditions show that 2AFC is the most efficient test, followed by Yes/No. Rating is not a good test for discrimination, but it does offer possibly valuable information on the naturalness of motion capture and model based motion separately, rather than just their discriminability. When significant observations were made by multiple test paradigms, their results agreed. While we have shown that the 2AFC test

is more efficient than a Yes/No test, there might be valid reasons to opt for a Yes/No test (see 1.2). In fact, for all results we obtained using both tests (that is, those dealing with only motion 1 and motion 2) the Yes/No test provided the same (significant) result as the 2AFC test did.

For procedural motion, generating a mix of coherent motion captured and procedurally generated motion to evaluate a procedural controller in isolation is more challenging. Perhaps one of the motion combination methods discussed in [1] can be used there.

Chaminade et al. [7] show that the sensitivity measure d' obtained from a 2AFC test that compared motion captured locomotion with key-framed locomotion is independent of the embodiment of a VH. If this result holds for other movement types and movement models, d' could prove an interesting measure to compare naturalness of motion models generated by different research groups.

Acknowledgments. This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie). We would like to thank Rob van der Lubbe for his help with SDT.

References

1. van Welbergen, H., van Basten, B.J.H., Egges, A., Ruttkay, Z., Overmars, M.H.: Real Time Animation of Virtual Humans: A Trade-off Between Naturalness and Control. In: Eurographics - State of the Art Reports, Eurographics Association, pp. 45–72 (2009)
2. Hartmann, B., Mancini, M., Pelachaud, C.: Formational parameters and adaptive prototype instantiation for mpeg-4 compliant gesture synthesis. In: Computer Animation, pp. 111–119. IEEE Computer Society, Los Alamitos (2002)
3. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents: Research articles. Comput. Animat. Virtual Worlds 15(1), 39–52 (2004)
4. Wooten, W.L., Hodgins, J.K.: Simulating leaping, tumbling, landing, and balancing humans. In: International Conference on Robotics and Animation, pp. 656–662 (2000)
5. van Welbergen, H., Zwiers, J., Ruttkay, Z.: Real-time animation using a mix of dynamics and kinematics. Submitted to Journal of Graphics Tools (2009)
6. Macmillan, N.A., Creelman, D.C.: Detection Theory: A User's Guide, 2nd edn. Lawrence Erlbaum, Mahwah (2004)
7. Chaminade, T., Hodgins, J.K., Kawato, M.: Anthropomorphism influences perception of computer-animated characters' actions. Social Cognitive and Affective Neuroscience 2(3), 206–216 (2007)
8. Weissenfeld, A., Liu, K., Ostermann, J.: Video-Realistic Image-based Eye Animation System. In: Eurographics 2009 - Short Papers, pp. 41–44. Eurographics Association (2009)
9. van Basten, B., Egges, A.: Evaluating distance metrics for animation blending. In: Proceedings of the 4th International Conference on Foundations of Digital Games, pp. 199–206. ACM, New York (2009)