© AAMAS This is the author's version of the work. It is posted here for your personal use. The definitive version was published in Jeffery Ventrella, Magy Seif El-Nasr, Bardia Aghabeigi, and Richard Overington. Gestural Turing Test: A Motion-Capture Experiment for Exploring Nonverbal Communication. AAMAS 2010 International Workshop on Interacting with ECAs as Virtual Characters, May 10, 2010. Retrieved from the Simon Fraser University Institutional Repository.

Gestural Turing Test

A Motion-Capture Experiment for Exploring Believability In Artificial Nonverbal Communication

Jeffrey Ventrella
Simon Fraser University, School of Interactive Art and Technology, Vancouver BC
jeffrey@ventrella.com

Magy Seif El-Nasr
Simon Fraser University, School of Interactive Art and Technology, Vancouver, BC
magy@sfu.ca

Bardia Aghabeigi
Simon Fraser University, School of Interactive Art and Technology, Vancouver BC
b.aghabeigi@gmail.com

Richard Overington
Emily Carr University of Art and Design, Vancouver, BC
roverington@ecuad.ca

ABSTRACT

One of the open problems in creating believable characters in computer games and collaborative virtual environments is simulating adaptive human-like motion. Classical artificial intelligence (AI) research places an emphasis on verbal language. In response to the limitations of classical AI, many researchers have turned their attention to embodied communication and situated intelligence. Inspired by *Gestural Theory*, which claims that speech emerged from visual, bodily gestures in primates, we implemented a variation of the Turing Test, using motion instead of text for messaging between agents. In doing this, we attempt to understand the qualities of motion that seem human-like to people. We designed two gestural AI algorithms that simulate or mimic communicative human motion using the positions of the head and the hands to determine three moving points as the signal. To run experiments, we implemented a networked-based architecture for a Vicon motion capture studio. Subjects were shown both artificial and human gestures, and were told to declare whether it was real or fake. Techniques such as simple gesture imitation were found to increase believability. While we require many such experiments to understand the perception of human-ness in movement, we believe this research is essential to developing a truly believable character.

Categories and Subject Descriptors

I.2.0 [Artificial Intelligence]: General
I.2.m [Artificial Intelligence]: Miscellaneous
J.4 [Social and Behavioral Sciences]: Psychology

General Terms

Algorithms, Measurement, Documentation, Performance, Design, Experimentation, Human Factors, Languages, Theory.

Keywords

Turing test, believability, gestural theory, virtual agent, motion capture, nonverbal, point light displays

1. INTRODUCTION

Alan Turing's thought-experiment of the 1950's was proposed as a way to test a machine's ability to demonstrate intelligent behavior. Turing had been exploring the question of whether machines can *think*. To avoid the difficulty of defining "intelligence", he proposed taking a behaviorist stance, and to ask: can machines do what we humans do? [19] In this thought experiment, a human observer engages in a conversation (using text-chat only) with two hidden agents – one of them is a real human and the other is an AI program. Both the human and the AI program try to appear convincingly human. If the observer believes that the AI program is a real human, then it passes the Turing Test.

The focus on verbal language in this and other explorations of intelligence is characteristic of classical AI research. Verbal language may indeed be the ultimate indicator of human intelligence, but it may not be the most representative indicator of intelligence in the broadest sense. Inventor/thinkers such as Rodney Brooks remind us that intelligence might be best understood, not as something based on a system of abstract symbols and logical decisions, but as something that emerges within an embodied, situated agent that must adapt within an environment [2]. If we can simulate at least some basic aspects of the embodied foundations of intelligence, we may be better prepared to then understand higher intelligence, and thus model and simulate believable behaviors in computer games and collaborative virtual environments. Justine Cassell said it well: "We need to locate intelligence, and this need poses problems for the invisible computer. The best example of located intelligence, of course, is the body." [3].

Gestural Theory [8] claims that speech emerged out of the more primal communicative energy of gesture. If this theory is correct, then perhaps we should explore this gestural energy as a viable indicator of intelligence. We have set up an experiment to run a Turing Test using an “alphabet” of three moving dots instead of the alphabet of text characters. In this experiment, the agents (one of then might be non-human) interact, and generate spontaneous body language through their ongoing interactions. One may ask: what is there to discuss if you only have a few points to wave around in the air? In the classic Turing Test, you can bring up any subject and discuss it endlessly. But remember the goal of the Turing Test: to fool a human subject into believing that an AI program is a human. However that is accomplished is up to the subject and the AI program. Turing chose the medium of text chat, which is devoid of any visual or audible queues. Body language was thus not an option for Turing. In contrast, we are using a small set of moving dots, and no verbal communication. Moving dots are abstracted visual elements (like the alphabet of written language), however, they are situated in time, and more intimately tied to the energy of natural language.

1.1 Prior Work, and Stated Contribution

Several variations of the Turing test based on simulated human motion have been implemented [20], [18], [11]. Imitation of human behavior has been shown to be effective in creating believability in virtual agents, such as work by Kipp [10] describing a system that uses imitation of human gesture to generate conversational gestures for animated embodied agents. Gorman [6] shows that imitation of the behaviors of a computer game player creates enhanced believability in artificial agents. Stone, et al [17] describe a technique for reproducing the structure of speech and gesture in new conversational contexts. Neff, et al, show how the gestural styles of individual speakers can be reconstructed, focusing on arm gestures [12]. The emotional and narrative content that emerges through extended interaction between a virtual agent and a human can be used for simulating memory and emotional states, thus increasing believability, as indicated by work by Seif El-Nasr [16]. Modeling the affective dimensions of characters and their personalities, as demonstrated by Gebhard [5], provides more robust, consistent behavior in an agent over extended time. While we have not developed such components, we have developed a scheme by which believability over extended interaction time can be measured.

Studies in using point light displays have shown that humans are sensitive to the perception of human movement, such as detecting human gait [1][9], and there are findings of distinct patterns of neural activity associated with the perception of human-made movement [15][14], indicating that a small number of visual elements can be used, not only for testing perception of believable motion, but also for use as control points in an animated character, using inverse kinematics (IK). IK is commonly used in computer animation to determine the joint rotations of a character, based on goal positions.

In our research we have chosen to reduce the visual aspect to a minimum, as a way to work with first principles of motion behavior and also to establish an efficient and manageable set of controllers for animating a character. This paper demonstrates a scheme for testing the believability in this highly-reduced set of primary motion features, using an established, well-studied method: the Turing Test.

We do not address the issue of *coverbal* gesture or ways to add a nonverbal layer to an existing verbal layer for conversational agents. This is basic research focusing on “silent copresence” and primitive communication through motion only.

1.2 Graphical Representation

The plastic human brain routinely adapts to new communication media and user interfaces. If a human communicator spends enough time “being” three dots, and communicating with three dots that behave in a similar way, then the body map quickly adapts to that schema. We have designed AI algorithms that “know” they exist as 3 dots, and must use a three-dot interface to communicate. Consider that an AI used for the classic Turing Test does not require the simulation of a mouth, tongue, lungs, diaphragm, or any apparatus used to generate verbal language, and nor does it require the simulation of fingers tapping on computer keyboards. Similarly, the Gestural Turing Test AI does not need to simulate the entire muscular and skeletal apparatus required for moving these points around. Semiosis is confined to the points only – they are the locus of communication.

How many points are needed to detect communicative motion? We had originally considered two dots – even *one* dot, as comprising the gestural alphabet. Our hypothesis was that, given enough time for a subject to interact with the dot(s), the intelligence behind it (or lack thereof) would eventually be revealed. With one dot, there would be very little indication of a physical human puppeteer – however, the existence of a human *mind* might become apparent over time, due to the spontaneous visual language that naturally would emerge, given enough time and interaction.

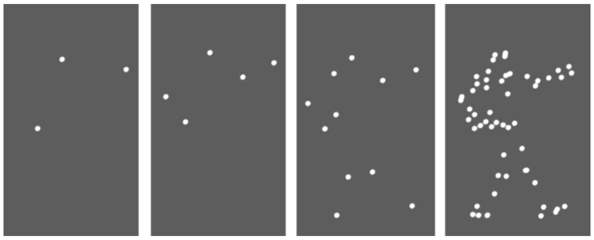

Figure 1. The human visual system can detect living motion from only a few dots.

Several points of light (dozens) would make it easier for the subject to discern between artificial and human (as illustrated in Figure 1). But this would require more sophisticated physical modeling of the human body, as well as a more sophisticated AI. For our experiment, we chose three points, because we believe the head and hands to be the most motion-expressive points of the body. The majority of gestural emblems originate in the head and hands.

2. EXPERIMENTS

Figure 2 shows a schematic of the studio setup. A human observer (the subject, shown at right) sits in a chair in front of a screen projected with two sets of three white dots.

The subject wears motion-capture markers (attached to a hat and two gloves), which are used to move the three dots on the right side of the screen. The three dots on the left side of the screen are moved by a hidden agent obscured by a room divider. This hidden agent is either another human with similar motion capture markers, or a software program that simulates human-made motion of the dots. No sounds or text can be exchanged between the subject and the hidden agent. Sign language is not possible, due to the limited number of dots.

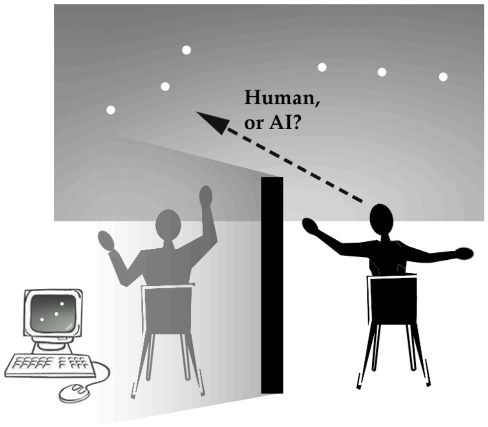

Figure 2. The human subject (right) interacts with the moving dots on the left and must decide if they are created by a hidden human or an AI program.

To generate the points from the human subjects we used the Vicon motion capture studio at Emily Carr University of Art and Design in Vancouver, BC. Figure 3 shows a screenshot of the Vicon interface (top). In order for the Vicon system to differentiate between the various objects in the scene, one of the hats used four markers, and each glove on the opposite side used two markers. Everything else used only one. This explains the linear-connected figures in the screenshot. This is only for purposes of calibration and disambiguation for the Vicon system, and makes no difference to the subject's view. Twenty cameras are deployed in the studio (six of them can be seen represented in wireframe at the top). Because of the room divider, some camera views of the markers are obscured, which accounts for occasional drop-out and flickering in the resulting points.

The stream of 3D positional data generated by the Vicon system while the subjects moved was distributed via a local area network to a laptop running the Unity game engine. We formatted the data using XML, which was also used for recording motions and archiving results from the experiments. A 3D scene consisting of six small white spheres – three on either side of a black divider – are animated by this data stream at 30 Hz. We used the Unity engine because we intend to extend this research to drive realistic avatars in a subsequent version of the project. An example display from Unity is shown at the bottom of Figure 3.

2.1 Artificial Gesture Algorithms

To generate the artificial gestures, we designed two algorithms. Both algorithms relied on the detection of the energy of the human's motions to trigger responsive gestures. We calculated energy from continually measuring the sum of the instantaneous speeds of the three points. If at any time energy changed from a value below a specified threshold to greater than that threshold, a response could be triggered, which depended on what kinds of gestures were playing at the time.

The first algorithm (AI1) employed a state-machine that chose among a set of short, pre-recorded gestures made by a human. These pre-recoded gestures included "ambient" motions (shifting in the chair, scratching, etc.) and a set of more dynamic gestures (emblems and communicative gestures such as waving, pointing, drawing shapes, "chair-dancing", etc.). When triggered, it played gestures form the set of dynamic gestures, and when it detected smaller movements, it responded by playing smaller, ambient gestures. AI1 did not include a sophisticated blending scheme for smooth transitions between gestures, and so it was often apparent to the subjects that it was not human. This was intentional: we wanted to expose the subjects to less-believable behaviors so that they could establish a base-level of non-believability on which to judge other motions.

The second algorithm (AI2) used a combination of procedurally-generated motions and imitative motions created while the experiment was being done. The procedurally-generated gestures were continuous (no explicit beginning, middle or end) and so any one of them could be blended in or out at any time, or layered together. These were constructed through combinations of several sine and cosine oscillations, with carefully-chosen phase offsets and frequencies. This included slight motions using a technique similar to Perlin Noise [13]. The imitative gestures were created by recording the positions of the subject's motions, translating them to the location where the AI was "sitting", and playing them back after about a second, with some variation, when a critical increase in energy was detected. AI2 used a more sophisticated blending technique, achieved by allowing multiple gestures (each with varying weights) to play simultaneously, such that the sum of the weights always equals 1. A cosine function was used for blending transitions to create an ease-in/ease-out effect, which helped to smooth transitions.

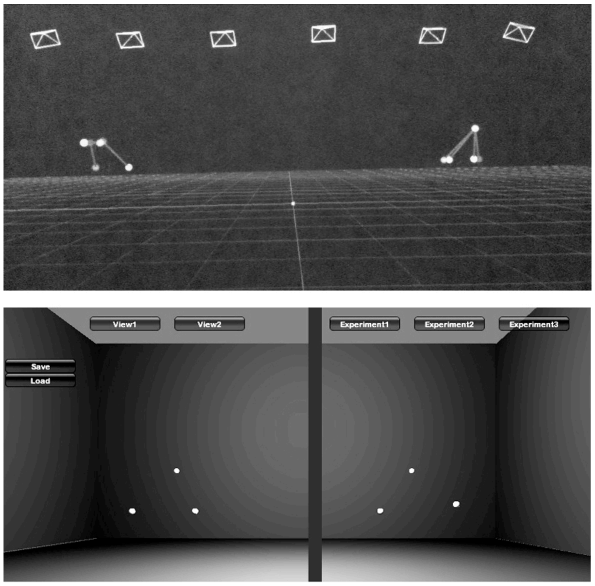

Figure 3. 'Skeleton template' of the Vicon motion capture system, required for labeling and calibration (top). Rendering with Unity engine, used in experiments (bottom)

Our rationale for recording the subject's gestures and playing them back, using AI2, was that it would not only appear human, but that it would also be imitative. Imitation is one of the most primary and universal aspects of communication – especially when there is a desire for rapport and emotional connection. Gratch, et. al, report that virtual agents that exhibit postural mirroring, and imitation of head gestures enhance the sense of rapport within subjects [7].

3. Results

There were 17 subjects. We ran 6 to 12 tests on each subject. A total of 168 tests were done. Figure 4 shows the results in chronological order from top to bottom. In this graph, the set of tests per subject is delineated by a gray horizontal line. The length of the line is proportional to the duration it took for the subject to make a response. The longest duration was just over 95 seconds. If the response was "false", a black dot is shown at the right end of the line. Wrong guesses are indicated by black rectangles at the right-side of the graph.

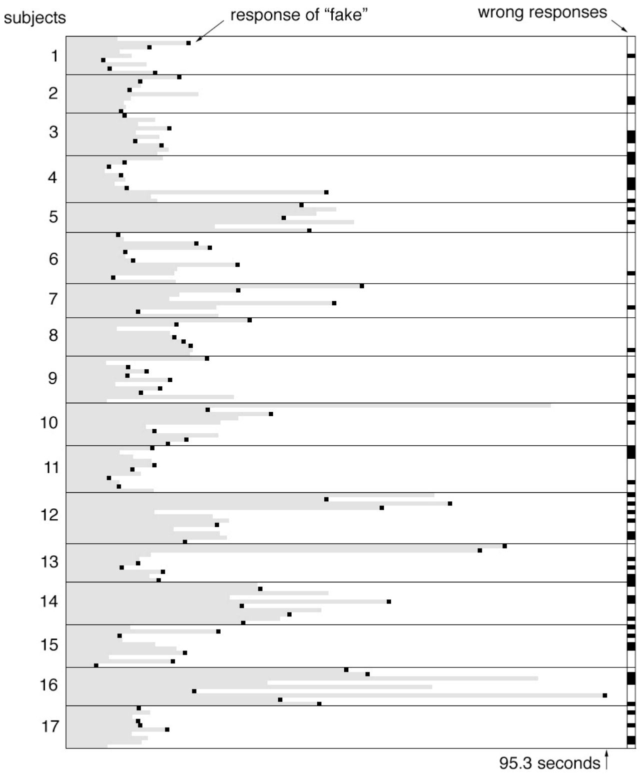

Figure 4. Test results displayed chronologically for all subjects

This graph reveals some differences in subjects' abilities to make correct guesses, and also differences in duration before subjects made a response. But we are more interested in how well the two AI algorithms performed against the human. This can be shown by separating out the tests according to which hidden agent was used (Human, AI1, or AI2). Figure 5 shows the percentages of wrong vs. right responses in the subjects for each of the three agents. As expected, the human had the most guesses of *real*. Also as expected, AI2 scored better than AI1 in terms of fooling subjects into thinking it was real.

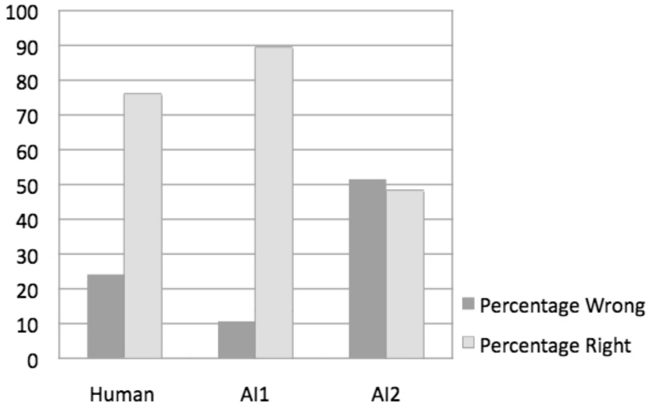

Figure 5. Percentages of right and wrong responses for each agent.

Believability in virtual agents can be measured in many ways – it doesn't have to be a binary choice, and in fact it has been suggested by critics of the classic Turing Test that its all-or-nothing test criterion may be a problem, and that a *graded* assessment might be more appropriate and practical [4]. One approach might be to measure how quickly a subject is convinced that a virtual agent is real. We calculated the average durations for each case of right and wrong responses, as shown in Table 1.

Table 1. Average durations for both right and wrong responses.

	Human		AI1		AI2	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Right	19.79	13.46	26.65	20.37	25.41	15.67
Wrong	23.38	19.42	24.75	26.96	22.72	15.79

The average durations before responding for the human agent are less than the average durations for the other agents, except for when subjects guessed wrongly for AI2. It may be that the authenticity of the human agent is easily and quickly determined, on average, which accounts for the slightly shorter average durations. But this is not conclusive. We ran a t-test and did not find any significant differences. In future experiments we would need more experimental data and more thorough analyses.

4. OBSERVATIONS

We selected both males and females, with the majority of the subjects being females. Some of the subjects displayed great confidence, and made quick decisions (which were not necessarily more correct). Some subjects took very long to decide (over a minute). We also found large variations among the kinds of gestures that the subjects made. Some of them were very reserved, holding their hands close together, and making small motions, while others made large motions. Some subjects gestured broadly and stopped to wait for a response, while others appeared to be swimming in place with no breaks or pauses. These variations in subject gesturing had a pronounced effect on AI response. Both AI algorithms relied on the gestures to be fast enough to trigger a response (specifically, the sum of the instantaneous speeds of the three points had to be greater than a certain threshold). Consequently, the subjects who made small, slow gestures were met with fairly uncommunicative artificial agents, and this had the compounding effect of less activity from the subject. It was not unlike two shy people who are unable to get a conversation going. This suggests to us that a more sophisticated AI would need to be designed that is able to gauge the overall energy of the subject's gestures and adjust its gesture-detection threshold accordingly.

4.1 Signal versus Noise

The Vicon motion capture system relies on multiple markers on the body to construct a reliable 3D representation. Since we used so few points, the system sometimes lost the labeling of points, and as a consequence, there was occasional drop-out and swapping between markers. This kind of problem is typically cleaned up in post-processing before motion capture data are used in a film, for instance. In our case, we were streaming the data in realtime, and had to manage this problem on the fly. At first we spent a bit of effort trying to remove these visual artifacts. But later we realized that the human subject may forgive the noise and still appreciate the signal, similar to the way that some static is tolerated in telephony. So, instead we decided to add *artificial artifacts* to the simulated output! The subject can easily discover that these glitches occur in his/her own two dots, and will quickly forgive them – seeing them in the artificial agent actually could actually enhance its believability. (Recall that in the Turing Test, *any* trick that can fool the human is fair game).

4.2 Problems

For subsequent tests, we would like to improve a number of things. For instance, the Vicon motion capture system is not set up to deal with very small numbers of markers. It relies on several markers, with many of them being a fixed distance apart, in order to keep track of the labeling of markers. Often, when the subject's hands were held together, the Vicon system would swap the hands, which required a recalibration in the middle of the test (this only took a second each time, but still we would prefer not to have to do it).

We cannot be 100% sure that a subject did not hear or "sense" that the human on the other side of the divider was the one gesturing. This is why we had originally planned on conducting the test in two remote locations (the rationale for developing a flexible network architecture). This setup however would require considerable technical work. Also, internet latency would introduce new problems that would have to be dealt with.

5. CONCLUSIONS

The results of these experiments show that when an artificial agent imitates the gestures of a human subject, there is more acceptance of that agent as being alive. We also show that this can be tested with a small set of visual indicators. This is consistent with research in studying human vision with point-light displays, and it suggests that believability need not be supported by visual realism.

The scope of this project did not permit the design of an extended AI with the ability to build on a collaborative semiotic process. But we believe that layering more sophisticated algorithms on top of the base behaviors we have implemented would create sustained believability over longer durations of time.

The contribution of the paper is twofold. First: the Gestural Turing Test itself can act as a methodology for validating gestural AI algorithms. It can easily be extended to include motion specified by many more control points, as well as any level of sophistication of intelligence, emotion, memory, natural language, and physical modeling. Secondly: the developed imitation algorithm can be analyzed to define a model or set of design lessons for creating better believable characters. The dots used in this experiment (which are actually projected 3D positions) are ultimately intended to become the control points for a fully-rendered avatar using IK. In a subsequent experiment, we intend to replace the graphical representation of dots with 3D avatars. It does not take a lot of control points to achieve reasonable motion, especially if the human model has a well-crafted constraints system to generate natural poses, given the pushing and pulling of the control points. One reason we feel that IK is a reasonable technique is because communicative motion often takes the form of hands moving in complicated paths, whereby the relative positioning of the hands (less so the elbows and shoulders) constitute the informational content. Thus, a gesturing system that is based on head and hand positioning (and rotation) over time is a valid scheme to use. One could even claim that communicative motion is IK-based at the neurological level. In future work, we aim to address several key questions related to the semantics of believable motions towards the development of techniques for designing believable characters.

6. ACKNOWLEDGMENTS

We would like to acknowledge the NSERC - NCE (Networks of Centres of Excellence of Canada), GRAND (graphics, animation and new media) for providing funds to support this project. We wish to thank Professor Leslie Bishko of the Emily Carr University of Art and Design for initial discussions and ideas, and for helping with access to the motion capture studios.

7. REFERENCES

[1] Blake, et al., Perception of Human Motion. 2007

[2] Brooks, Rodney A. Elephants Don't Play Chess. Robotics and Autonomous Systems 6, 1990 3-15

[3] Cassell, J. "Embodied Converstational Agents: Representation and Intelligence in User Interfaces," *AI Magazine*, vol. 22, 2001.

[4] French, R. If it walks like a duck and quacks like a duck... The Turing Test, Intelligence and Consciousness. Oxford Companion to Consciousness. Wilken, Bayne, Cleeremans (eds.) Oxford Univ. Press. 461-463. 2009.

[5] Gebhard, P. ALMA – a Layered Model of Affect. Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems. 29-36

[6] Gorman, B., Thurau, C., Bauckhage, C., and Humphrys, M. Believability Testing and Bayesian Imitation in Interactive Computer Games. From Animals to Animats 9. Springer, 2006.

[7] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. Creating Rapport with Virtual Agents. Proceedings of the 7th international conference on Intelligent Virtual Agents. Springer, 2007.

[8] Hewes, Gordon W. 1973. Primate Communication and the Gestural Origins of Language. Current Anthropology 14:5-24.

[9] Jacobs, A., and Pinto, J. Experience, Context, and the Visual Perception of Human Movement. Journal of Experimental Psychology. 2004, Vol. 30, No. 5, 822–835

[10] Kipp, M. Gesture Generation by Imitation. Dissertation for Doctor of Engineering. Saarland University, Saarbruecken, Germany. Published by Dissertation.com. 2003

[11] Livingstone. Turing's Test and Believable AI in Games. 2006

[12] Neff, M. et al. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics. Volume 27, Issue 1. 2008

[13] Perlin, K. Real Time Responsive Animation with Personality. IEEE Transactions on Visualization and Computer Graphics. Volume 1, Issue 1, March, 1995

[14] Pinto, J., and Shiffrar, M. The visual perception of human and animal motion in point-light displays. Social NeuroScience Volume 4, Issue 4 August 2009 , pages 332 – 346

[15] Saygin, A.P., Wilson, S., Hagler, D., Bates, E., and Sereno, M. I.· Point-Light Biological Motion Perception Activates Human Premotor Cortex. The Journal of Neuroscience, July 7, 2004.

[16] M. Seif El-Nasr, T. Ioerger, and J. Yen, "FLAME - Fuzzy Logic Adaptive Model of Emotions," *Autonomous Agents and Multi-Agent Systems*, vol. 3, pp. 219-257, 2000.

[17] Stone, M., et al. Speaking with Hands, Speaking with Hands: Creating Animated Conversational Characters from Recordings of Human Performance. ACM Transactions on Graphics Volume 23, Issue 3. 2004

[18] Stuart, J. et al. Generating Novel, Stylistically Consonant Variations on Human Movement Sequences. 2007

[19] Turing, Alan. Computing Machinery and Intelligence, 1950

[20] Van Welbergen. Informed Use of Motion Synthesis Models. Motion in Games. Springer, 2008.