The Saliency of Anomalies in Animated Human Characters

JESSICA HODGINS, Carnegie Mellon University & Disney Research, Pittsburgh
SOPHIE JÖRG, CAROL O'SULLIVAN, Trinity College Dublin
SANG IL PARK, Sejong University
MOSHE MAHLER, Carnegie Mellon University

Virtual characters are much in demand for animated movies, games, and other applications. Rapid advances in performance capture and advanced rendering techniques have allowed the movie industry in particular to create characters that appear very human-like. However, with these new capabilities has come the realization that such characters are yet not quite "right." One possible hypothesis is that these virtual humans fall into an "Uncanny Valley", where the viewer's emotional response is repulsion or rejection, rather than the empathy or emotional engagement that their creators had hoped for. To explore these issues, we created three animated vignettes of an arguing couple with detailed motion for the face, eyes, hair, and body. In a set of perceptual experiments, we explore the relative importance of different anomalies using two different methods: a questionnaire to determine the emotional response to the full-length vignettes, with and without facial motion and audio; and a 2AFC (two alternative forced choice) task to compare the performance of a virtual "actor" in short clips (extracts from the vignettes) depicting a range of different facial and body anomalies. We found that the facial anomalies are particularly salient, even when very significant body animation anomalies are present.

Categories and Subject Descriptors: I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Animation*

General Terms: Experimentation, Human Factors

Additional Key Words and Phrases: Human animation, virtual characters, perception of human motion, motion capture, eye tracking

ACM Reference Format:

Hodgins, J., Jörg, S., O'Sullivan, C., Park, S., and Mahler, M. 2010. The saliency of anomalies in animated human characters. ACM Trans. Appl. Percept. 7, 4, Article 22 (July 2010), 14 pages.
DOI = 10.1145/1823738.1823740 http://doi.acm.org/10.1145/1823738.1823740

1. INTRODUCTION

Recent advances in computer animation have allowed the creation of very lifelike virtual characters for movies such as *Polar Express, Beowulf, A Christmas Carol* (Robert Zemeckis) and *Avatar* (James Cameron). Numerous articles in the public press and research domains (e.g., Geller [2008]) have speculated that the greater realism of these characters occasionally falls into the Uncanny Valley

This research was supported in part by Disney Research Pittsburgh, the Irish Research Council for Science Engineering and Technology (IRCSET), and NSF CCF-0811450.

Author's addresses: J. Hodgins, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213; S. Jörg, C. O'Sullivan (corresponding author). Trinity College Dublin, College Green, Dublin 2, Ireland; email: carol.osullivan@cs.tcd.ie; S. Park, Sejong University, 98 Gunja-Dong, Gwangjin-Gu, Seoul, 143-747 Republic of Korea; M. Mahler, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.

DOI 10.1145/1823738.1823740 http://doi.acm.org/10.1145/1823738.1823740

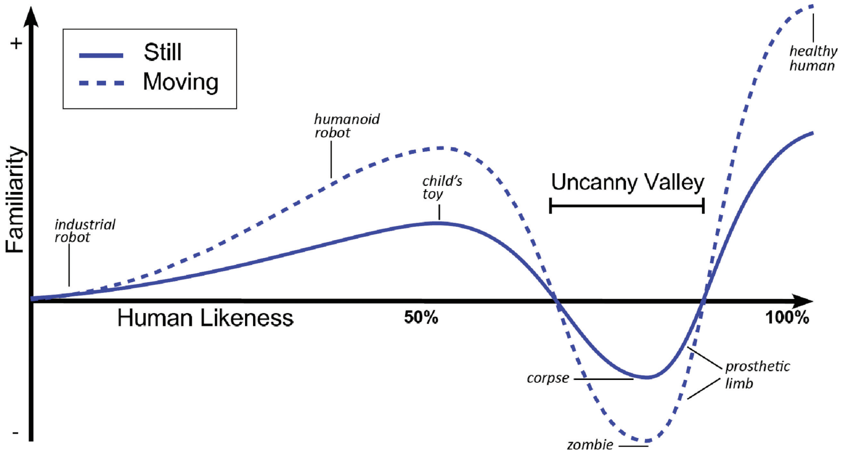

Fig. 1. The hypothesized graph for the Uncanny Valley response (redrawn from Mori [1970]).

hypothesized by Mori [1970]. When this happens, audiences tend to feel a sense of eeriness or revulsion rather than empathizing with the characters or becoming emotionally engaged in the story.

Mori hypothesized two valleys, a shallower one for appearance and a deeper valley (or stronger effect) for motion, as seen in Figure 1. For example, a humanoid robot with some anthropomorphic features may appear familiar enough to be appealing, but different enough from a real person so as not to cause fear. On the other hand, a highly realistic android (or indeed a wax-work model) may look so lifelike that it might be convincing at first glance. However, when it moves in a not quite human-like manner, it can remind people of the living dead, and hence may fall to the bottom of the hypothesized valley. A parallel in computer graphics could be the contrast between audience responses to the appealing, cartoon-like characters in Pixar's *The Incredibles* and the more human-like, yet somehow off-putting humans in *Polar Express* [Levi 2004]. Many different causes for this supposed revulsion have been put forward, including lack of familiarity, the appearance of disease or death, and a mismatch between elements of the character design and the motion. Similarly, appearance or motion artifacts have been suggested as the cause; stiff or "wooden" faces, incorrect eye gaze, and stiff body motion are all frequently proposed.

However, the concept is far from precise and it is not clear what the axes of the graph should be. Indeed, it is unlikely that viewers' complex, multisensory responses to virtual characters can be described by such a simple graph [Tinwell and Grimshaw 2009]. Nevertheless, it provides a useful conceptual framework within which to explore these important issues. Given our limited understanding of this phenomenon, we need techniques to analyze the impact of the improvements and remaining anomalies in human motion and appearance on the perception and interpretation of a scene. Our aim is to capture the higher-level cognitive and emotional responses that the complex scenarios found in movies or games are intended to elicit. Simple psychophysical measures, such as perceptibility thresholds, are not sufficient to characterize participants' responses to such scenarios. Therefore, we attempt to capture higher-level reactions, such as the level of engagement with the characters, or what kind of emotion the participants felt when viewing them, for example, did they sympathize with them, did they think they were convincing in the scene, or did they just not feel any emotion at all when viewing them?

In this article, we examine the impact of anomalies using two different experimental methods: one using "vignettes," that is, animated scenes of about 30 seconds in length that tell a story with emotional content; and the other using shorter "clips," depicting 3–5-second extracts from the vignettes. We measure the response to the vignettes by asking questions that attempt to capture participants'

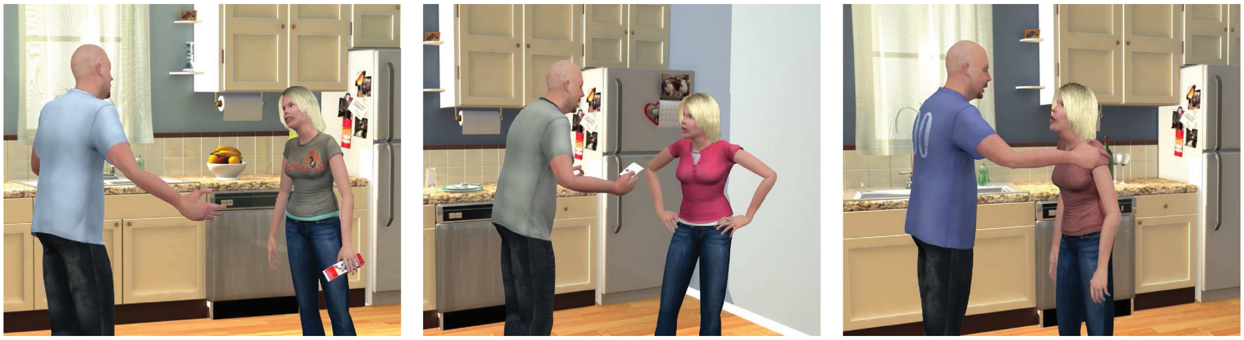

Fig. 2. Scenes from three vignettes portraying arguments between a man and a woman over a shortage of milk, unwisely spent money, and moving out when the relationship is over. These examples include facial, eye, whole body, and hair motion.

emotional engagement and sympathy for the characters, such as: “How angry was the woman?” and “How justified (appropriate) were the man’s reactions?” In the experiments with short clips, participants had the simpler task of choosing one of a pair. However, we provide context for the experiments: they were “auditioning” a virtual actress and were asked to choose in which clip her performance was more compelling. We pitched different face and body anomalies directly against each other, requiring participants to make the choice of “In which of these two clips was the virtual actress more convincing?” Again, this response is at a higher level than simply looking explicitly for anomalies, for example, by asking in which clip they saw an animation artifact.

The three vignettes depict an angry reaction to a shortage of milk, unwisely spent money, and moving out at the end of a relationship. Modern motion capture systems and carefully planned capture sessions allow for the creation of very lifelike virtual characters. For our full animation, we included animation of the face of the characters, eye tracking data for the character facing the camera (i.e., the woman), simulated hair motion for the female character, as well as full body motion for both characters (Figure 2). However, for complex scenarios it may sometimes be necessary to sacrifice some elements of the animation because of the exigencies of the capture process. For example, capturing eye or face motion simultaneously with whole body motion is complicated due to the higher resolution required for the smaller, more detailed regions and the effect of body-worn capture devices on the actor’s performance. Some elements of a scene such as clothing and hair may be impossible to capture and must be added later via simulation.

Using the shorter clips, we explored two different classes of stimuli: those that we felt represented common flaws in animation (e.g., eyes not tracked, only simple facial motion captured); and those that represented disease conditions (Bell’s Palsy, tremor, immobile shoulder joint). With the former, we aimed to provide insights into what should be added to a standard capture to create a more effective character; the latter was designed to explore a question raised by MacDorman [2005] in the context of the Uncanny Valley of whether the eeriness is caused by our natural revulsion when presented with animations that resemble disease conditions. We hoped that by exploring these two classes of stimuli, we could gain insight into these practical and scientific questions simultaneously.

We found that facial anomalies were most salient in the short clips, especially when they resembled medical conditions such as Bell’s Palsy, demonstrating that this is where the most effort in animation should be directed. In the long vignettes, the interaction between the presence or absence of sound and/or facial animation affected the responses to questions about emotional impact, indicating that any attempt to model the Uncanny Valley must take multiple dimensions into account. These results point to some interesting issues to be explored in future work.

2. RELATED WORK

Although the goal of most movies is to have an emotional impact on the audience, there is little work exploring the perceptual consequences of errors or degradation in motion for scenes with affective content. A number of researchers have evaluated the effect of motion manipulations on participants' perception of human animations. In most of these studies, multiple short views of a single character walking, jumping, or dancing were shown. User responses were typically used to derive perceptual thresholds or sensitivity measures. Hodgins and colleagues [1998] demonstrated that changes in the geometric models used for rendering affected participants' sensitivity to small changes in human motion. Wang and Bodenheimer [2004] ran a user study to determine the optimal blend length for transitions between two motion clips. McDonnell and colleagues [2007] investigated the relationship between frame rate and smooth human motion using a series of psychophysical experiments. Reitsma and Pollard [2003] demonstrated that velocity and acceleration errors introduced in the motion editing process are noticeable for ballistic motions. Finally, Harrison and colleagues [2004] used a simple link structure to examine how the limbs of more realistic models may be stretched without the viewer perceiving the distortion. The major difference between these studies and our experiments is that we are using vignettes with significant emotional content (anger) and attempt to measure how participants' emotional response is affected by significant changes to the character's motion.

The Uncanny Valley has been directly explored most often in the context of robots because that was the setting for Mori's original hypothesis in 1970. These studies are limited in that they most often use images of existing robots or virtual characters and as such do not span the space of designs but just evaluate the best efforts of skilled robot or character designers [Bartneck et al. 2007; Schneider et al. 2007; MacDorman et al. 2009]. To address this concern, researchers have also performed morphs between two existing designs (or between a robot or virtual character and a photograph of a similarly posed human). Some of these studies have produced a curve resembling that of the Uncanny Valley although the question remains whether this result arises from an imperfect morph or from a true instance of the Uncanny Valley. For example, double images of the eyes sometimes arise in a morph between two images that are not sufficiently well aligned. Hanson and colleagues [2005] performed a carefully designed morph between a cartoon rendering of a female character and a similarly attired and posed image of a woman. They found no Uncanny Valley in a Web-based survey of the "acceptability" of the images.

Because of the difficulty of developing controllable stimuli, the existence of the Uncanny Valley for motion has been tested in only a few experiments [Ho et al. 2008; MacDorman 2006]. Using videos of 17 robots ranging from wheeled soccer playing robots to androids and one human video, Ho et al. [2008] explored various possible axes for the Uncanny Valley graph, including eerie, creepy, and strange. Correlations between perceived eeriness, voice and lip synchronization, and facial expression have also been found, where the lack of synchronization was particularly disturbing [Tinwell and Grimshaw 2004]. We hope that by creating our vignettes from motion capture data, we can provide greater control over the conditions in our experiments and thereby provide a framework that will support more rigorous evaluation.

Researchers have also explored different measures for assessing participants' responses to images or video suspected to span the Uncanny Valley. MacDorman developed an elegant experiment using terror management techniques as his metric [MacDorman 2005]. In this study, he found that a photo of an android elicited a heightened preference for worldview supporters and a diminished preference for worldview threats relative to a control group that viewed a photo of a woman. These results indicate that perhaps an "eerie" robot elicits an innate fear of death or disease (as measured by the terror

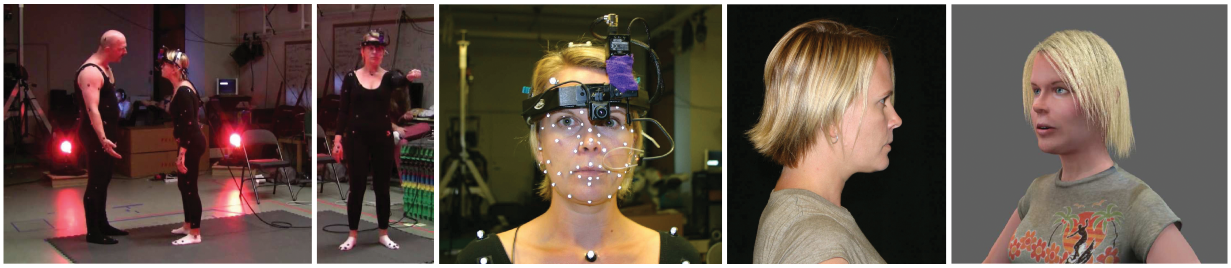

Fig. 3. (a) capture session; (b) calibration of the eye tracker by looking at a known location (the hand); (c) the eye tracker as worn by the female actor; (d) the hair of the female actor, and (e) the rendered hair.

management metrics). Patterns of gaze fixation have also been used to determine whether an android is responded to in a similar manner as a human or a nonhumanoid robot [MacDorman et al. 2005].

3. EXPERIMENT DESIGN

Our goal is to determine how degradation of human motion affects the emotional response of participants to an animation. Our hypotheses are that: *facial anomalies* will have the most effect (as McDonnell et al. [2009] and others have shown that people focus most on the faces of real and virtual humans); Conditions relating to *human diseases* will have a similarly strong effect (as postulated by MacDorman [2005]); The presence or absence of *audio* will influence responses, as it is to be expected that much of the emotional content of a vignette is delivered through this medium.

To test these hypotheses, we generated three vignettes, each of which depicted an event with emotional content (see Figure 2). Two professional actors (one male, one female) performed based on our instructions and plot outlines. All three vignettes take place in the kitchen of a house or apartment shared by a couple and last approximately 30 seconds. For the first vignette, *Milk*, the female actor is angry about the fact that there was no milk for her breakfast cereal. For the second vignette, *Money*, the male actor is angry because he just picked up the mail and found a large bill for clothing. In the final vignette, *MovingOut*, the couple has split up and the man finds that the woman has a guest when he returns unexpectedly to pick up his possessions.

3.1 Creation of the Stimuli

An optical motion capture system consisting of 18 Vicon cameras was used to record the positions of 286 retro-reflective markers attached to the bodies, fingers, and faces of two actors (Figure 3(a)). The data for the hand and body markers was mapped onto a skeleton. Joint angles and root motion were computed from this data and used to drive character models constructed in Maya. Photographs were used as reference material to ensure that the character models resembled the captured actors.

To capture the facial expression of the actors, we placed about 50 reflective small markers on each actor's face, after which our facial animation pipeline deformed the detailed facial geometry to match the movement of the markers. We followed the method of Park and Hodgins [2006]: the markers are first segmented into seven near-rigid parts (Figure 4) such that markers belonging to a part are considered to move roughly together. The rigid motion of each part is approximated with six degrees-of-freedom (three for translation and three for rotation) from the movement of its member markers. The remaining local deformation of a part is modeled with quadratic deformations and radial basis deformations, then added to the rigid body motion. This method provides a good representation of the actors' facial movements. One difference between our method and those used for emotion capture in the recent movie, *Avatar*, is that their facial motion passes through the hands of skilled animators, while ours

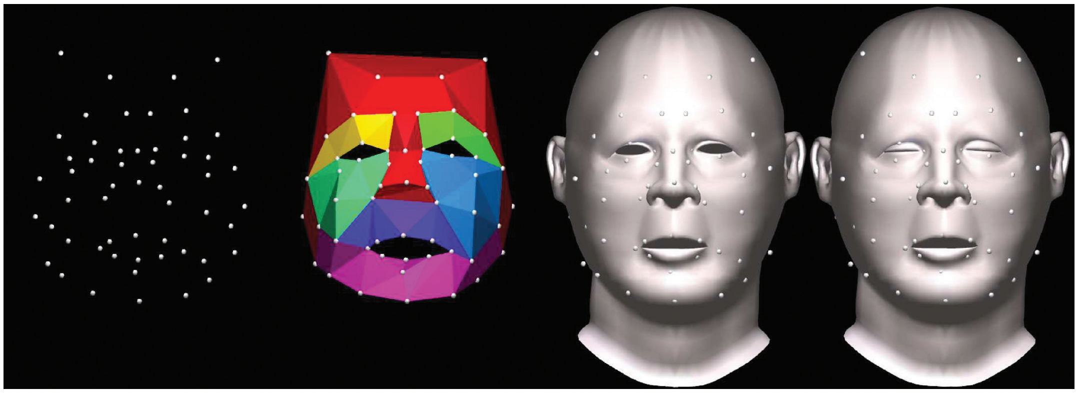

Fig. 4. (a) facial markers; (b) seven near-rigid parts; (c) eye-open facial geometry; (d) eye-closed facial geometry.

was automatically processed. Their facial motion is captured with a wearable camera rig (required for a large capture region) while ours was captured with fixed cameras. It should, of course, be noted that they applied the captured motion to nonhuman characters, whereas our focus is on realistic human characters.

The blinking movement of the eyelid can also be captured from the optical markers. However, because there is not much space for multiple markers around the eyelids, the resulting facial geometry with only one marker attached to each eyelid is not convincing when the eyes are closed. We created a separate facial geometry with closed eyes (see Figure 4) and blended it linearly with the resulting eye-open geometry when a blink occurred. The timing of the blinking is extracted from the data, using the distance between two markers placed on the upper and lower eyelids, respectively, to detect the frames at which they start to close, are fully closed, start to open, and are fully open.

The hair motion was created with Maya 2008. After generating an initial state for the hair, the position/velocities of the head motion and gravity are then used to drive the simulation. We selected simulation parameters that were intended to make it resemble the fine blond hair of our female actor (Figure 3(d)). The man's shirt was also simulated in Maya 2008.

The eye movements of the female actor were recorded using a head-mounted ASL eye-tracker which measured the direction of the gaze of the actor's left eye at 120fps (Figure 3(c)). The eye tracker was calibrated by gazing at a known location in the motion capture space (Figure 3(b)). This data was used for the motion of both the left and right eyes in the final rendering. The resulting characters were placed into a model of a kitchen and rendered. Finally, we recorded the voices of the actors in a separate recording session, which is common practice in games and movies.

Four different types of *Long* vignette stimuli were created by displaying all three unabridged vignettes using one of the following four experimental conditions: FullFace with Sound; FullFace with NoSound; NoFace with Sound; and NoFace with NoSound. For the NoFace conditions, we rendered the animations with no facial or eye movements, while for NoSound we turned off the audio. To create a second set of *Short* clip stimuli, we divided the vignettes up into shorter snippets that still made sense, for example, a full sentence, each between 3–6 seconds long. By modifying the original animation, we created six facial and three body anomalies for the woman character and multiple clips were selected for each condition. This set of stimuli was sufficient to allow random selection in the experiment and thus reduce repetition. The *Short* clip experimental conditions were as follows.

—Full (F): the unmodified animation

—No Face (NF): the face and eyes were completely frozen

—No Eyes (NE): the face was animated but the eyes did not move

—Offset Eyes (OE): one eye was rotated by a constant amount, giving an effect similar to Amblyopia (or lazy eye)

—Low Face, Eyes (LF_E): the motion of the jaw was restricted to one degree of freedom (open/close for speech synchronization), but the eyes were animated

—Low Face, No Eyes (LF_NE): face as with LF_E, but the eyes were not animated

—Half Face (HF): only half the face was animated—this condition was chosen to mimic Bell's palsy

—No Arm (NA): one arm was not animated and was held stiffly in front of the character as if in a sling

—Noisy Arm (NyA): a tremor was added to the character's arm

—No Pelvis (NP): the top half of the body only was animated by removing data from the root and legs. The motion of the two spine joints was smoothed to reduce artifacts caused by the lack of motion in the pelvis

3.2 Method

In order to examine the effect of degraded human motion on participants' perception of emotional virtual characters, we conducted a series of experiments. First, to confirm the results of McDonnell et al. [2009] and others that faces attract the most attention, we tracked the eye motion of nine participants while they watched the *Long* vignettes, using a Tobii eye-tracker. Participants were told that they would be asked questions about the vignettes afterwards, but we did not analyze this data. While we know that the task being performed can greatly affect eye movements, the knowledge that one will be asked general questions is a reasonably neutral task. Therefore, we can assume that this task will give a good indication of where attention is focused in general. As this pretest was intended as a simple confirmation of whether participants did indeed tend to focus on faces more when viewing the vignettes under normal conditions, we only examined the eye movement data of participants for the FullFace with Sound vignettes. Furthermore, in order to elicit as unbiased a response as possible, we only examined the eye movements of each participant while watching the first vignette (as there may have been effects of familiarity with the scene during subsequent viewings, thereby introducing bias). Therefore, we examined the data of nine participants, with three watching either Milk, Money, or MovingOut first, and then averaged over the three participants for each vignette. We found that they focused on the faces 54%, 79%, and 68% of the time (i.e., the percentage of overall fixation durations) in the Milk, Money, and Moving Out vignettes, respectively.

For the *Long* vignettes and *Short* clips experiments, we recruited small groups of participants by advertising on university email lists, posters, and fliers. This recruitment effort resulted in a varied participant pool drawn from a range of different disciplines and backgrounds: 48 males and 37 females ranging in age from 18–45, with 76% of them reporting passing knowledge or less with 3D computer graphics; all gave their informed consent. There were 85 individual participants, of whom 69 did both the *Long* vignettes and *Short* clips experiments; the remaining 16 only participated in a follow-up *Short* clip experiment at the end. Both experiments took approximately 20 minutes each. A total of 28 small groups participated, ranging in size between one and six participants. The experiments were all held in the same room, where the vignettes and clips were projected onto a 150 × 80 cm screen at a distance of 2–2.5m to the participants, giving a horizontal/vertical field of view of around 35 × 20 degrees. We asked participants not to make audible responses to the animations and they were not permitted to discuss their opinions with each other during the study. An experimenter was present at all times to ensure compliance. At the start of the experiment, the participants filled out some general information such as gender and age.

Those who participated in both experiments first watched all three *Long* vignettes (in the order Milk, Money, MovingOut) displayed in one of the four conditions: that is, with FullFace or NoFace, and with Sound or NoSound. They recorded their answers on a questionnaire sheet. These questions required them to rate, on a 5-point scale, the *anger* levels of both characters, how *justified* their reactions were, and how much *sympathy* they had for them. They were also asked to select who was mainly responsible for the argument and to write some free-form text about the events. They then watched a random ordering of the *Short* clips after the scenario was set as follows:

> We have captured the voice and motions of the female actor and now want to know which virtual characteristics are best at conveying a compelling performance. Therefore, you are "auditioning" a virtual actor. You will be asked after viewing each pair to indicate your answer to the following question: "In which clip was the virtual actor more convincing."

The participants indicated their choice for each pair of clips using a 2AFC (two alternative forced-choice) answer sheet. Each group viewed a random selection of four clips for every combination of conditions, as follows: 35 participants looked at all combinations of pairs of the six different FaceType clips and the full animation, for example, they watched Full versus NoEyes (F/NE), LowFace Eyes versus LowFace NoEyes (LF_E/LF_NE) and all other possible pairs; a further 34 participants viewed FaceType/Body combinations depicting the five disease conditions and the full animation, for instance, Full versus NoArm (F/NA), NoisyArm versus OffsetEyes (NyA/OE), NoPelvis versus HalfFace (NP/HF). We had avoided using the NoFace condition in the clips as we felt that this would be too obvious. However, as a follow-up to determine whether this hypothesis was indeed correct, the final 16 participants only performed a *Short* clips experiment, where they looked at all paired combinations of the set of clips including NoFace, the Full animation and the most obvious of the other artifacts (i.e., NoArm, OffsetEyes, and HalfFace, as determined by the first set of experiments).

In these experiments, we wanted to create large effects that were certain to elicit differentiated responses. Therefore, we aimed to generate clearly suprathreshold stimuli, which were well above the Just Noticeable Difference for the conditions tested. We did informal pretests with people who were familiar with the animations and tuned the magnitude to what was definitely noticeable without being too extreme. We are confident that all stimuli were approximately equally noticeable when viewed directly by an expert.

4. RESULTS

Statistical analysis was carried out on all results in order to test the significance of the effects found. To analyze the responses to each question in the *Long* vignettes experiment, we conducted a 3-factor, repeated measures Analysis of Variance (ANOVA), where the between groups factors were *AudioType* (i.e., Sound, NoSound) and *FaceType* (i.e., FullFace, NoFace) and the within groups factor was *Vignette* (i.e., Milk, Money, MovingOut). Post hoc analysis was performed on significant effects using Newman-Keuls tests for pairwise comparison of means. The most significant results are shown in Figure 5 and a summary of significant effects (i.e., with $p < 0.05$) is shown in Table I. The most significant results were found for the woman, which was expected as her face was visible throughout the three vignettes. There was a significant positive correlation between the ratings for the sympathy and justified ratings (Spearman's $\rho = 0.75, p < 0.05$). Therefore, we do not include these results, or those relating to questions about the man, as they provide no added insights. There was also a significant correlation between sympathy and anger ratings, but this was much lower ($\rho = 0.26, p < 0.05$) and different effects were observed for this question.

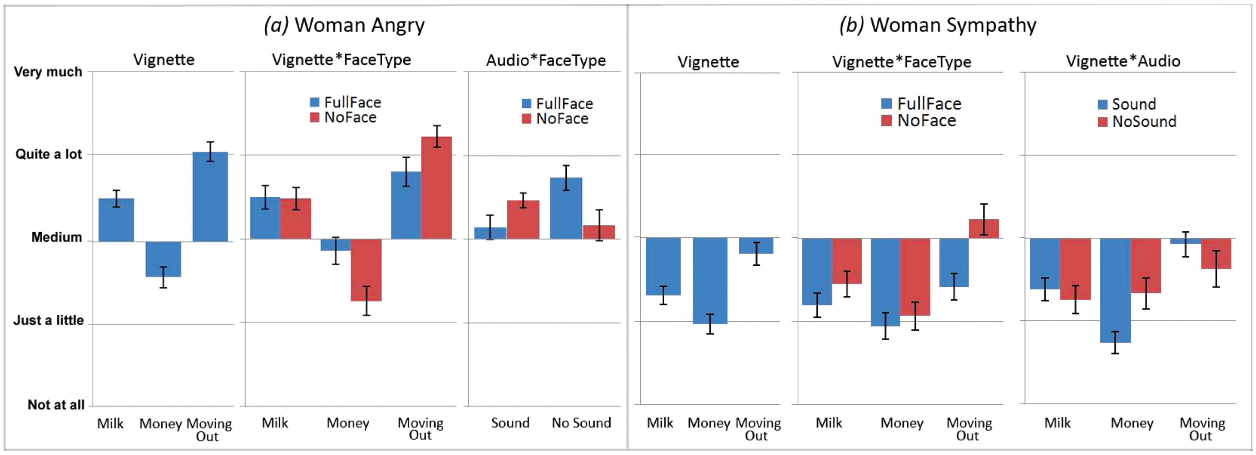

Fig. 5. Most interesting results for the *Long* vignettes questions

Table I. Significant Results for the *Long* Vignettes Experiments ($p < 0.05$ in all Cases)

Question	Effect	F-Test	Post-hoc
Woman Angry	Audio∗FaceType	$F(1, 65) = 9.8401$	FullFace NoSound > FullFace Sound, FullFace NoSound > NoFace NoSound
	Vignette	$F(2, 130) = 61.229$	Money < Milk < MovingOut
	Vignette∗FaceType	$F(2, 130) = 7.7607$	Money FullFace different to everything Money NoFace different to everything MovingOut FullFace different to everything MovingOut NoFace different to everything
Woman Sympathy	FaceType	$F(1, 65) = 5.1201$	FullFace<NoFace
	Vignette	$F(2, 130) = 16.923$	Money < Milk < MovingOut
	Vignette∗Audio	$F(2, 130) = 6.0725$	Money WithSound < all
	Vignette∗FaceType	$F(2, 130) = 4.1188$	MovingOut NoFace > all
Man Angry	Vignette	$F(2, 130) = 44.479$	Milk < MovingOut < Money
Man Sympathy	Vignette	$F(2, 130) = 6.4253$	Money > all

In the *Long* vignettes, as expected, a main effect of *Scene* was found for all questions. In particular, the Milk, Money, and MovingOut vignettes were found to elicit distinct levels of anger and sympathy ratings for the woman (see Figure 5 and Table I). An *Audio**FaceType* interaction for the Angry question (Figure 5(a)) revealed that she was angriest in the FullFace NoSound condition and much less so when Sound was present. This result suggests that while her facial animation did appear to express anger, the sound appears to have diluted the emotion, perhaps because we recorded the soundtrack separately. A *Vignette**FaceType* interaction shows that she was most angry in the MovingOut vignette with NoFace, and least angry in the Money vignette with NoFace. Perhaps this result can be explained by our eye-movement results, which showed that the face was most fixated upon in the Money vignette and less so in the MovingOut vignette. Hence, the lack of facial animation in Money reduced the overall level of expression, whereas in the MovingOut vignette, the lack of facial expression distracted less from the more physical body interactions between the couple. Or it may have been the case that her "stony" face simply came across as extremely angry in that vignette.

For the Sympathy question (Figure 5(b)), a main effect of *FaceType* is explained by the fact that the woman was found to be significantly less sympathetic when her face was animated (FullFace) than when it was not (NoFace). The reasons for this may be more apparent from examining the interactions.

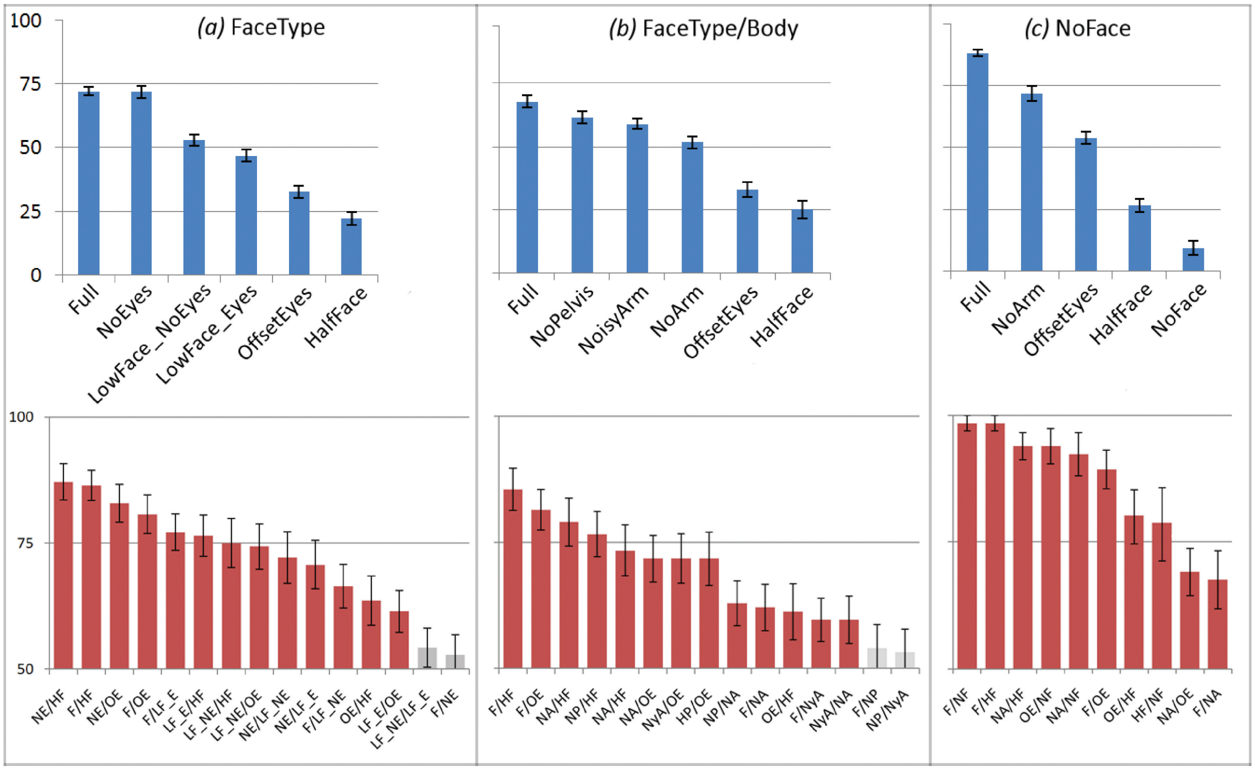

Fig. 6. All results for the *Short* clips comparison experiments: overall (top) and pairwise (bottom) preferences; vertical axis shows % times preferred and error bars show one standard error.

It should be noted that, for this question, medium sympathy can be considered to be the "neutral" level, with ratings above or below the bar representing positive or negative reactions. No conditions elicited a positive rating towards the woman, but negative reactions are also important to engage viewers (e.g., an evil villain). Therefore, the lower sympathy rating with the facial animation actually implies a stronger negative reaction to the woman. A *Vignette*FaceType* interaction occurred because she was rated most sympathetic in the MovingOut vignette with NoFace. However, this result was closest to neutral and therefore implies a lack of expressivity. The animated face results imply a stronger negative reaction to her, perhaps because her anger came across as too strong for the circumstances and hence people disliked her more. This was reflected in some of the comments from the FullFace conditions, such as "The woman seemed more aggressive and louder"; "The man was more explaining than arguing"; "Physically aggressive woman." The *Vignette*Audio* interaction was caused by the woman being least sympathetic in the Money vignette with sound.

These results reinforce the common wisdom that facial animation is a key component of expression and careful attention should be paid to how the face is used to convey emotion. However, how the audio is matched or dubbed to the facial animation is also clearly key, as it interacted with the other factors in nonobvious ways.

The results for the *Short* clips experiments are shown in Figure 6. To compare the means of the overall preferences for the FaceType, FaceType/Body, and NoFace sets of clip conditions (shown on the top), we carried out a single factor, repeated measures ANOVA for each. Post hoc analysis was

Table II. Significant Differences for Overall *Short* Clips Preferences ($p < .05$ in all cases)

FACETYPE	FACETYPE/BODY	NOFACE
Main Effect: $F(5, 170) = 66.57$	*Main Effect:* $F(5, 150) = 33.24$	*Main Effect:* $F(4, 60) = 131.38$
Full > LowFace_NoEyes, LowFace_Eyes	Full > NoArm, OffsetEyes, Half Face	All significantly different
Full > OffsetEyes, HalfFace	NoPelvis > NoArm, OffsetEyes, HalfFace	
NoEyes > LowFace_NoEyes, LowFace_Eyes	NoisyArm > OffsetEyes, HalfFace	
NoEyes > OffsetEyes, HalfFace	NoArm > OffsetEyes, HalfFace	
LowFace_NoEyes > OffsetEyes, HalfFace		
LowFace_Eyes > OffsetEyes, HalfFace		
OffsetEyes > HalfFace		

performed using Newman-Keuls tests and the significant effects are reported in Table II. For the pairwise comparisons (Figure 6, bottom), single t-tests were carried out to determine if they were real preferences (i.e., significantly different from 50%), and those that were not are shown on the graph in grey.

These results were very informative because we were able to run more conditions and use a within-subjects design. For the 35 (18M,17F) participants who viewed the FaceType combinations of clips (i.e., all face conditions pitched against each other), as expected the *Full* animation was preferred overall, except to *NoEyes* (Figure 6(a)). It was surprising that the lack of eye movements was not more disturbing, but perhaps it is because we included eye-blinks, even though the eyes themselves were frozen. As hypothesized, the two facial "disease" conditions were found to be the worst, with OffsetEyes being slightly but significantly more preferred to HalfFace. With the 34 (24M, 10F) participants that compared these two conditions with body anomalies (FaceType/Body), they found the facial anomalies significantly more disturbing than any of the (quite significant) body errors (Figure 6(b)). In fact, some participants reported not noticing any body anomalies at all, as they were completely focussed on the faces. HalfFace and OffsetEyes were found to be equally disturbing, with the next least preferred being NoArm. In a follow-up with 16 (10M, 6F) participants, we tested the effect of NoFace, to determine whether it really was the worst condition (Figure 6(c)): as the results of our *Long* vignette experiments showed that emotional content was conveyed even in the absence of facial motion, we postulated that the two worst facial anomalies may be found to be more disturbing. However, we found that the NoFace condition was significantly less preferred, even to the worst of the other anomalies.

5. DISCUSSION

We have developed a new type of experimental framework that did, as planned, elicit higher-level responses from participants to ecologically valid stimuli (similar to those that would be found in real applications). Now, we and others can build on this work to run new experiments to provide further guidelines and insights for animators and the developers of animation systems. These promising approaches to investigating these issues are applicable for evaluating other types of problems, not just motion capture.

In summary, we found that removing facial animation and/or sound from the *Long* vignettes did change the emotional content that was communicated to our participants. The woman's behavior was found to be particularly affected by these four conditions, as she was the most visible in the scenes. Clearly, even the animations with no sound and no facial animation conveyed emotional content, but less effectively than when facial animation was included (making the woman more unsympathetic, for example).

The results from the *Short* clips experiments were unambiguous. Facial anomalies are more disturbing than quite significant body motion errors, so this is where the most effort should be expended to achieve natural human motion. While we expected that facial anomalies would dominate over body motion errors, it was more significant than we had predicted. Furthermore, we could also derive some more nuanced guidelines regarding the significance of different types of facial anomalies relative to one another. An interesting observation is that although the NoFace condition in the *Long* vignettes conveyed emotional information to the participants (even when sound was absent), in the clips it was almost never preferred to any of the other conditions. However, in the written comments after viewing the vignettes, several participants referred to the absence of facial motion and expressed annoyance. This observation is worthy of further investigation, as are the effects of other types of emotion, for example, happy or sad, or more neutral.

What do these results tell us about the Uncanny Valley theory? As we predicted, the two face disease conditions and the stiff arm were preferred least, therefore supporting a hypothesis that these motions fell into the Uncanny Valley because they reminded people of illness or injury. On the other hand, the facial anomalies were much more disturbing than the body motion errors, which shows that the valley may be affected by attention. The frozen face was preferred least of all in the clips, perhaps because it resembles a corpse, yet the woman in the MovingOut vignette with no facial animation was rated as being most angry. This could have been caused by her appearing "stony-faced", which is an extreme form of anger. Alternatively, it may have been caused by focussing more attention on the quite aggressive body language of the couple, again pointing to a possible effect of attention. More extensive eye-tracking studies would be very useful here to determine if this is indeed the case. Audio was also an important factor when watching the vignettes, as it interacted with both the vignette and the facial animation in different ways. This effect was most evident for the Money vignette, where the absence of sound actually increased the perceived anger of the woman. All of these results point to a complex, multidimensional model of uncanniness which is unlikely to be a valley but rather a parameterizable space.

We intentionally designed our stimuli to be suprathreshold. However, without calibrating the different magnitude changes we applied to our stimuli, it is difficult to rank order them in the perceptual effect that they have (i.e., one could argue that it might be that one change was "bigger" than another rather than that faces are more important than arms). Therefore, psychophysical studies to explore Just Noticeable Differences (JND) would provide further useful insights. We could, for example, run JND experiments, fit an Ogive function to the data, and then run our experiments on, say, the stimuli at 75% perceptibility. Now that we have an effective framework for eliciting responses from participants, and some indicative results, we can explore how to probe around these extremes in even more targeted ways to expose greater detail about the responses and to appropriately model these data.

Further study is needed to explore these effects more deeply. Some participants seemed to suggest that there was a mismatch in emotion between the body and the voice for the Full animation with Sound conditions. For example: one participant wrote: "Both have unnaturally calm voices given their phyical actions", while another commented: "Both were really angry (supposingly!)". As it is common in practice to capture motion and audio tracks separately (e.g., for performance capture with a VIP or A-List actor, or when animated movies are dubbed in another language), if the emotional content is being disrupted by the bodies being more or less emotional than the voices, or desynchronized due to bad dubbing, this is worthy of further investigation (e.g., see Giorgolo and Verstraten [2008]; Carter et al. [2010]).

With the *Long* vignette experiments, the types of questions we asked showed that higher-level responses can be elicited from natural, ecologically valid stimuli. We were able to run only a limited set of our conditions as each subject was shown only one modification (on the three vignettes) and we had to use an across-subjects design. Now that we have established a framework for exploring these factors with longer, more natural stimuli, we can consider developing more detailed or sophisticated questionnaires to explore them further across a broader range of stimuli. Our framework can now be extended to investigate, for example, differences in responses to cartoon rendering, video, and realistic rendering. With longer vignettes, we could consider using "sliders" that allow people to record their emotions over time, and not just at the end of a vignette [Lottridge 2008]. Physiological measures such as heart-rate [Meehan et al. 2002], brain responses [Anders et al. 2004], and more in-depth eye-movement analysis [Hunt et al. 2007] could also provide insight. Finally, an interesting question is whether the same effects are found when anomalies occur in less familiar, nonhuman characters, as in *Avatar*.

ACKNOWLEDGMENTS

We would like to thank the members of the graphics lab at Carnegie Mellon University for their assistance with creating the motion data, and all participants in our experiments.

REFERENCES

ANDERS, S., LOTZE, M., ERB, M., GRODD, W., AND BIRBAUMER, N. 2004. Brain activity underlying emotional valence and arousal: A response-related fMRI study. *Hum. Brain Map. 23*, 200–209.

BARTNECK, C., KANDA, T., ISHIGURO, H., AND HAGITA, N. 2007. Is the uncanny valley an uncanny cliff? In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'07)*. IEEE, 368–373.

CARTER, E. J., SHARAN, L., TRUTOIU, L., MATTHEWS, I., AND HODGINS, J. 2010. Perceptually motivated guidelines for voice synchronization in film. *ACM Trans. Appl. Percept. 7*, 4.

GELLER, T. 2008. Overcoming the uncanny valley. *IEEE Comput. Graph. Appl. 28*, 4, 11–17.

GIORGOLO, G. AND VERSTRATEN, F. 2008. Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*. 31–36.

HANSON, D., OLNEY, A., PRILLIMAN, S., MATHEWS, E., ZIELKE, M., HAMMONS, D., FERNANDEZ, R., AND STEPHANOU, H. 2005. Upending the uncanny valley. In *Proceedings of the National Conference on Artificial Intelligence (AAI'05)*.

HARRISON, J., RENSINK, R. A., AND VAN DE PANNE, M. 2004. Obscuring length changes during animated motion. *ACM Trans. Graph. 23*, 3, 569–573.

HO, C., MACDORMAN, K., AND DWI PRAMONO, Z. 2008. Human emotion and the uncanny valley: A glm, mds, and isomap analysis of robot video ratings. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI'08)*. 169–176.

HODGINS, J. K., O'BRIEN, J. F., AND TUMBLIN, J. 1998. Perception of human motion with different geometric models. *IEEE Trans. Visual. Comput. Graph. 4*, 4, 307–316.

HUNT, A. R., COOPER, R. M., HUNGR, C., AND KINGSTONE, A. 2007. The effect of emotional faces on eye movements and attention. *Vis. Cogn. 15*, 513–531.

LEVI, S. 2004. Why Tom Hanks is less than human; While sensors cannot capture how humans act, humans can give life to digital characters. *Newsweek 650*, 305–306.

LOTTRIDGE, D. 2008. Emotional response as a measure of human performance. In *ACM Extended Abstracts on Human Factors in Computing Systems (CHI'08)*. 2617–2620.

MACDORMAN, K. 2005. Mortality salience and the uncanny valley. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. 399–405.

MACDORMAN, K. 2006. Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *Proceedings of the ICCS/CogSci. Long Symposium: Toward Social Mechanisms of Android Science*.

MACDORMAN, K., GREEN, R., HO, C., AND KOCH, C. 2009. Too real for comfort? Uncanny responses to computer generated faces. *Comput. Hum. Behav. 25*, 3, 695–710.

MACDORMAN, K., MINATO, T., SHIMADA, M., ITAKURA, S., COWLEY, S., AND ISHIGURO, H. 2005. Assessing human likeness by eye contact in an android test bed. In *Proceedings of the XXVII Annual Meeting of the Cognitive Science Society*.

MCDONNELL, R., LARKIN, M., HERNÁNDEZ, B., RUDOMIN, I., AND O'SULLIVAN, C. 2009. Eye-Catching crowds: Saliency based selective variation. *ACM Trans. Graph. 28*, 3, 1–10.

MCDONNELL, R., NEWELL, F., AND O'SULLIVAN, C. 2007. Smooth movers: Perceptually guided human motion simulation. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 259–269.

MEEHAN, M., INSKO, B., WHITTON, M., AND BROOKS, JR., F. P. 2002. Physiological measures of presence in stressful virtual environments. *ACM Trans. Graph. 21*, 3, 645–652.

MORI, M. 1970. Bukimi no tani (the uncanny valley). *Energy 7*, 4, 33–35. (Translated from the Japanese by K. F. MacDorman and T. Minato).

PARK, S. I. AND HODGINS, J. K. 2006. Capturing and animating skin deformation in human motion. *ACM Trans. Graph. 25*, 3, 881–889.

REITSMA, P. AND POLLARD, N. 2003. Perceptual metrics for character animation: sensitivity to errors in ballistic motion. *ACM Trans. Graph. 22*, 3, 537–542.

SCHNEIDER, E., WANG, Y., AND YANG, S. 2007. Exploring the uncanny valley with japanese video game characters. In *Situated Play*, B. Akira, Ed. 546–549.

TINWELL, A. AND GRIMSHAW, M. 2004. Survival horror games - An uncanny modality. In *Thinking After Dark International Conference*.

TINWELL, A. AND GRIMSHAW, M. 2009. Bridging the uncanny: An impossible traverse? In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era (MindTrek'09)*. ACM, New York, 66–73.

WANG, J. AND BODENHEIMER, B. 2004. Computing the duration of motion transitions: An empirical approach. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 335–344.

Received June 2010; accepted June 2010