ECAG 2008 Workshop

Facial and Bodily Expressions for Control and Adaptation of Games

CTIT PROCEEDINGS OF THE WORKSHOP ON
FACIAL AND BODILY EXPRESSIONS FOR
CONTROL AND ADAPTATION OF GAMES (ECAG'08)

Amsterdam, the Netherlands, September 16, 2008

Anton Nijholt and Ronald Poppe (eds.)

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Nijholt, A., Poppe, R.W.

Facial and Bodily Expressions for Control and Adaptation of Games (ECAG 2008)
Proceedings of the Workshop on Bodily Expressions for Control and Adaptation of Games
A. Nijholt, R.W. Poppe (eds.)
Enschede, Universiteit Twente, Faculteit Elektrotechniek, Wiskunde en Informatica

ISSN 1568–7805
CTIT Workshop Proceedings Series WP08-03

trefwoorden: facial expression, body movement, voluntary control, involuntary control,
games, adaptation of games, exertion interface, user adaptation

Book orders:
Charlotte Bijron
University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
P.O. Box 217
NL 7500 AE Enschede
tel: +31 53 4893740
fax: +31 53 4893503
Email: bijron@cs.utwente.nl

Druk- en bindwerk: PrintPartners Ipskamp, Enschede

Facial and Bodily Expressions for Control and Adaptation of Games (ECAG'08)

September 16, 2008, Amsterdam
Workshop organized in conjunction with the
2008 IEEE International Conference on Automatic Face and Gesture Recognition
(September 17-19, http://www.fg2008.nl)

Programme Chairs and Organizers

Anton Nijholt (HMI, University of Twente, the Netherlands)
Ronald Poppe (HMI, University of Twente, the Netherlands)

Program Committee

Jeremy Bailenson (Stanford University, USA)
Nadia Berthouze (University College London, UK)
Antonio Camurri (Universty of Genova, Italy)
Yun Fu (University of Illinois at Urbana-Champaign, USA)
Hatice Gunes (University of Technology, Sydney, Australia)
Mitsuru Ishizuka (University of Tokyo, Japan)
Nadia Magnenat-Thalmann (University of Geneva, Switzerland)
Christopher Peters (Université de Paris 8, France)
Mannes Poel (University of Twente, the Netherlands)
Gang Qian (Arizona State University, USA)
Rainer Stiefelhagen (Universität Karlsruhe, Germany)

Additional Reviewers
Betsy van Dijk, Dirk Heylen, Zsofi Ruttkay, and Wim Fikkert (all University of Twente, the Netherlands)
Ioannis Patras (Queen Mary, University of London)

Technical Proceedings Editor
Hendri Hondorp

Copies

Copies of these proceedings can be ordered from:
HMI Secretariat: hmi_secr@cs.utwente.nl
University of Twente

August 18, 2008, Anton Nijholt, Ronald Poppe (eds)

Contents

Preface: Facial and Bodily Expressions for Control and Adaptation of Games

Anton Nijholt and Ronald Poppe
University of Twente, Dept. of Computer Science, Human Media Interaction Group
P.O. Box 217, 7500 AE Enschede, The Netherlands
{a.nijholt,poppe}@ewi.utwente.nl

1. Ambient Intelligence Environments

In future Ambient Intelligence (AmI) environments, we assume intelligence embedded in the environment, in its devices (furniture, mobile robots) and in its virtual human-like interaction possibilities. These environments support the human inhabitants and visitors of these environments in their activities and interactions by perceiving them through their sensors (e.g. cameras, microphones). Support can be reactive but also, and more importantly, pro-active and unobtrusive, anticipating the needs of the inhabitants and visitors by sensing their behavioral signals and being aware of the context in which they act. Health, recreation, sports and playing games are among the needs inhabitants and visitors of smart environments will have. Sensors in these environments can detect and interpret nonverbal activity and can give multimedia feedback to invite, stimulate, advise and engage. Activity can aim at improving physical and mental health, but also at improving capabilities related to a profession (e.g. ballet), recreation (e.g. juggling), or sports (e.g. fencing). Plain fun, to be achieved from interaction, can be another aim of such environments.

Such AmI environments know about the user. Maybe, rather than talk about a user, we should talk about an inhabitant, a gamer, a partner or an opponent. Humans will partner with such environments and their devices, including virtual and physical human-like devices (physical robots and virtual humans). Sensors and display technologies allow us to design environments and devices that offer implicit, explicit and human-like interaction possibilities. In particular, these environments allow multimodal interaction with mixed and augmented virtual reality environments, where these environments know about human interaction modalities and also know about how humans communicate with each other in face-to-face, multi-party, or human-computer interaction. Knowing about the 'user' means also that the environment knows about the particular 'user'. Indeed, smart environments identify users, know about their context and know about their preferences. Dealing with preferences and anticipating user behavior requires collecting and understanding patterns of user behavior.

Sensors embedded in current and future AmI environments allow reactive and pro-active communication with inhabitants of these environments. The environment, its devices and its sensors can track users, can recognize and anticipate the actions of the user and can, at least that is our assumption, interpret facial expressions, head movements, body postures and gestures that accompany turn-taking and other multi-party interaction behavior. There is still a long way to go from nowadays computing experiences to future visions where we can experience interactions in mixed reality and virtual worlds, integrated in smart sensor-equipped physical environments, and allowing seamless perceptual coherence when we have our body and our interactions mediated between the real and the virtual worlds and vice versa. Nevertheless, there are already applications where we have interactive systems observing the body movements and facial expressions of a human inhabitant or user of a particular environment and use information obtained from such observations to guide and interpret a user's activities and his interactions with the environment [1–4].

2. Ambient Entertainment Environments

The video game market is still growing. But there is also the success of the dance pads of Dance Dance Revolutions, Nintendo's Wii and its applications for games, sports and exercises, and Sony's EyeToy. Rather than using keyboard, mouse or joystick, there are sensors that make a game or sports application aware of a gamer's activities. The application can be designed in such a way that the gamer consciously controls the game by his activities (e.g., using gestures or body movements to navigate his avatar in a 3D game environment or to have a sword fight with an enemy avatar). The application can also use the information that is obtained from its sensors to adapt the environment to the user (e.g., noticing that the gamer needs more challenges).

We mentioned 3D environments and avatars. There are many applications (sports, games, leisure, and social communication) where we want to see ourselves acting and performing in virtual worlds and where we want to have others seeing us acting and performing in these virtual worlds.

We may want our nonverbal expressions displayed on our avatar in social communication. We may want our moods and emotions expressed by our avatar in a game or in a Second Life-like environment. This allows us to increase our presence in these environments and it allows others present and represented in these environments to communicate with us in natural, human-like, ways. It requires the sensors to mediate our, often unconsciously displayed, non-verbal social cues in the interaction with virtual game environments. It also requires sensors to mediate our consciously produced gestures, facial expressions, body postures, and body movements that are meant to have effect on the environment or on its synthesized virtual inhabitants.

3. Control and Adaptation of Games: The Workshop

In this workshop of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008), the emphasis is on research on facial and bodily expressions for the control and adaptation of games. We distinguish between two forms of expressions, depending on whether the user has the initiative and consciously uses his or her movements and expressions to control the interface, or whether the application takes the initiative to adapt itself to the affective state of the user as it can be interpreted from the user's expressive behavior. Hence, we look at:

- Voluntary control: The user consciously produces facial expressions, head movements or body gestures to control a game. This includes commands that allow navigation in the game environment or that allow movements of avatars or changes in their appearances (e.g. showing similar facial expressions on the avatar's face, transforming body gestures to emotion-related or to emotion-guided activities). Since the expressions and movements are made consciously, they do not necessarily reflect the (affective) state of the gamer.

- Involuntary control: The game environment detects, and gives an interpretation to the gamer's spontaneous facial expression and body pose and uses it to adapt the game to the supposed affective state of the gamer. This adaptation can affect the appearance of the game environment, the interaction modalities, the experience and engagement, the narrative and the strategy that is followed by the game or the game actors.

The workshop shows the broad range of applications that address the topic. For example, Cockburn *et al*. present a game where obstacles can be avoided by performing facial expressions. The game is used to help children with Autism Spectrum Disorder to improve their facial expression production skills. Bodily control is used to play a quiz in libraries, presented by Speelman and Kröse. Children can answer questions by pointing at answers, and dragging choices around. The experience of the game was compared to mouse control. A further investigation of the type of body movements that users make when playing Wii games is done by Pasch *et al*. They analyze motion capture data and user observations to identify different playing styles.

Van den Hoogen *et al*. present a study into the involuntary behavior of users that play video games. They measure mouse pressure and body posture shifts, which they correlate to the user's arousal level. Wang and Marsella present a game that invokes emotion in the user, and investigate the variety of observed facial expressions. Work by Wiratanaya and Lyons regards both voluntary and involuntary control. By reacting on a user's involuntary behavior, the user is encouraged to engage in a conscious interaction with a virtual character. They intend to use their work to entertain and engage dementia sufferers. Wang *et al*. present a system that reproduces observed facial expressions in an efficient manner, to be used in online applications. This type of work can be used in combination with automatic facial animation systems such as presented by Orvalho.

At ECAG, invited talks were given by Louis-Philippe Morency on "Understanding Nonverbal Behaviors: The Role of Context in Recognition", and by Nadia Bianchi-Berthouze on the experience of interacting with physically challenging games. We are grateful to the program committee, the FG 2008 organization and all others that helped in organizing this workshop.

References

[1] A. T. Larssen, T. Robertson, L. Loke, and J. Edwards. Introduction to the special issue on movement-based interaction. *Journal Personal and Ubiquitous Computing*, 11(8):607–608, November 2007.

[2] F. Müller, S. Agamanolis, M. R. Gibbs, and F. Vetere. Remote impact: Shadowboxing over a distance. In M. Czerwinski, A. Lund, and D. Tan, editors, *Extended Abstracts on Human Factors in Computing Systems (CHI'08)*, pages 2291–2296, Florence, Italy, April 2008.

[3] A. Nijholt, D. Reidsma, H. van Welbergen, R. op den Akker, and Z. Ruttkay. Mutually coordinated anticipatory multimodal interaction. In A. Esposito, N. Bourbakis, N. Avouris, and I. Hatzilygeroudis, editors, *Nonverbal Features of Human-Human and Human-Machine Interaction*, volume 5042 of *Lecture Notes in Computer Science (LNCS)*, pages 73–93, Patras, Greece, October 2008.

[4] D. Reidsma, H. van Welbergen, R. Poppe, P. Bos, and A. Nijholt. Towards bi-directional dancing interaction. In R. Harper, M. Rauterberg, and M. Combetto, editors, *Proceedings of the International Conference on Entertainment Computing (ICEC'06)*, volume 4161 of *Lecture Notes in Computer Science (LNCS)*, pages 1–12, Cambridge, United Kingdom, September 2006.

SmileMaze: A Tutoring System in Real-Time Facial Expression Perception and Production for Children with Autism Spectrum Disorder

Jeffrey Cockburn[1], Marian Bartlett[2], James Tanaka[1],
Javier Movellan[2] , Matthew Pierce[1] and Robert Schultz[3]

[1]Department of Psychology, University of Victoria, Victoria, British Columbia V8W 3P5 Canada

[2] Institute for Neural Computation, University of California, San Diego, La Jolla, CA 92093-0445, USA

[3] Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA

Abstract

Children with Autism Spectrum Disorders (ASD) are impaired in their ability to produce and perceive dynamic facial expressions [1]. The goal of SmileMaze is to improve the expression production skills of children with ASD in a dynamic and engaging game format. The Computer Expression Recognition Toolbox (CERT) is the heart of the SmileMaze game. CERT automatically detects frontal faces in a standard web-cam video stream and codes each frame in real-time with respect to 37 continuous action dimensions [2]. In the following we discuss how the inclusion of real-time expression recognition can not only improve the efficacy of an existing intervention program, Let's Face It!, but it allows us to investigate critical questions that could not be explored otherwise.

1. Introduction

The field of computer vision has made significant progress in the past decade, notably within the domain of automated facial expression recognition. The field has now matured to a point where its technologies are being applied to important issues in behavioral science. Cutting edge computer vision technologies can now be leveraged in the investigation of issues such as the facial expression recognition and production deficits common to children with *autism spectrum disorder* (ASD). Not only can these technologies assist in quantifying these deficits, but they can also be used as part of interventions aimed at reducing deficit severity.

The Machine Perception Lab at University of California, San Diego, has developed the *Computer Expression Recognition Toolbox* (CERT), which is capable of measuring basic facial expressions in real-time. At the University of Victoria in British Columbia, the *Let's Face It!* (LFI!) system was developed as a training program that has been shown to improve the face processing abilities of children with ASD. By combining the expertise behind these two technologies in disparate disciplines we have created a novel face expertise training prototype, *SmileMaze*. SmileMaze integrates the use of facial expression production into an intervention program aimed at improving the facial expression recognition skills of children with ASD. In the following text we will describe CERT, LFI!, and their union, SmileMaze, a novel facial expression recognition and production training prototype. We also wish to pay special attention to the scientific opportunities beyond the technologies themselves. In particular, we will discuss how an interdisciplinary approach combining cutting edge science from both computer and behavioral sciences has provided an opportunity to investigate high impact issues that were previously intractable.

2. CERT: The Computer Expression Recognition Toolbox

Recent advances in computer vision open new avenues for computer assisted intervention programs that target critical skills for social interaction, including the timing, morphology and dynamics of facial expressions. The Machine Perception Laboratory at UCSD has developed the *Computer Expression Recognition Toolbox* (CERT), which analyzes facial expressions in real-time. CERT is based on 15 years experience in automated facial expression recognition [3] and achieves unmatched performance in real-time at video frame rates [4]. The

system automatically detects frontal faces in a video stream and codes each frame with respect to 37 continuous dimensions, including basic expressions of anger, disgust, fear, joy, sadness, surprise, as well as 30 facial action units (AU's) from the Facial Action Coding System.

The technical approach is a texture-based discriminative method. Such approaches have proven highly robust and fast for face detection and tracking [5]. Face detection and detection of internal facial features is first performed on each frame using boosting techniques in a generative framework [6]. Enhancements to Viola and Jones include employing Gentleboost instead of AdaBoost, smart feature search, and a novel cascade training procedure, combined in a generative framework. Automatically located faces are rescaled to 96x96 pixels, with a typical distance of roughly 48 pixels between the centers of the eyes. Faces are then aligned using a fast least squares fit on the detected features, and passed through a bank of Gabor filters with 8 orientations and 9 spatial frequencies (2:32 pixels per cycle at 1/2 octave steps). Output magnitudes are then normalized and passed to the facial action classifiers.

Facial action detectors were developed by training separate support vector machines to detect the presence or absence of each facial action. The training set consisted of over 8000 images from both posed and spontaneous expressions, which were coded for facial actions from the Facial Action Coding System. The datasets used were the Cohn-Kanade DFAT-504 dataset [7]; The Ekman, Hager dataset of directed facial actions [8]; A subset of 50 videos from 20 subjects from the MMI database [9]; and three spontaneous expression datasets collected by Mark Frank (D005, D006, D007) [10]. Performances on a benchmark datasets (Cohn-Kanade) show state of the art performance for both recognition of basic emotions (98% correct detection for 1 vs. all, and 93% correct for 7 alternative forced choice), and for recognizing facial actions from the Facial Action Coding System (mean .93 area under the ROC over 8 facial actions, equivalent to percent correct on a 2-alernative forced choice).

In previouse experiments, CERT was used to extract new information from spontanious expressions [11]. These experiments addressed automated discrimination of posed from genuine expressions of pain, and automated detection of driver drowsiness. The analysis revealed information about facial behavior during these conditions that were previously unknown, including the coupling of movements. Automated classifiers were able to differentiate real from fake pain significantly better than naïve human subjects, and to detect driver drowsiness above 98% accuracy. Another experiment showed that facial expression was able to predict perceived difficulty of a video lecture and preferred presentation speed [12]. Statistical pattern recognition on large quantities of video data can reveal emergent behavioral patterns that previously would have required hundreds of coding hours by human experts, and would be unattainable by the non-expert. Moreover, automated facial expression analysis enables investigation into facial expression dynamics that were previously intractable by human coding because of the time required to code intensity changes.

3. LFI!: The Let's Face It! program

While most people are social experts in their ability to decode facial information, an accumulating body of evidence indicates that individuals with *autism spectrum disorder* (ASD) lack many of the rudimentary skills necessary for successful face communication. ASD is clinically diagnosed as impaired socialization and communicative abilities in the presence of restricted patterns of behavior and interests [13].

3.1. Facial recognition deficits in ASD

Children with ASD frequently fail to respond differentially to faces over non-face objects, are impaired in their ability to recognize facial identity and expression, and are unable to interpret the social meaning of facial cues. For children with ASD, facial identity recognition is specifically impaired in the midst of a normally functioning visual system [14]. Also, children with ASD demonstrate marked impairment in their ability to correctly recognize and label facial expressions [15].

Recognizing faces, identification of expression, and recognition of identity are fundamental face processing abilities. However, the pragmatics of everyday face processing demand that people go beyond the surface information of a face in an effort to understand the underlying message of its sender. For example, in the real world, we read a person's eye gaze to decipher what they might be thinking, or we evaluate a person's expression to deduce what they might be feeling. Not surprisingly, children with ASD also show deficits in eye contact [16], joint attention [17], and using facial cues in a social context [18].

3.2. Let's Face It!: A computer-based intervention for developing face expertise

Let's Face It! (LFI!) is a computer-based curriculum designed to teach basic face processing skills to children with ASD [19]. For ASD populations, there are several advantages to a computer-based approach. Children with

ASD may actually benefit more from computer-based instruction than traditional methods [20]. Computer- versus teacher-based approaches in object naming skills have also been compared [21]. It was found that children in the computer-based instruction learned significantly more new words and showed greater motivation for learning activity than children in the traditional teacher-based approach. Also, the features, such as music, variable-tone intensity, character vocalizations, and dynamic animations, are particularly motivating and reinforcing for persons with ASD and can easily be incorporated into computer-based instruction [22]. Finally, a computer-based curriculum offers a way to provide cost-effective instruction to ASD children in either a home or school setting.

LFI! targets skills involved in the recognition of identity, interpretation of facial expressions and attention to eye gaze through a set of diagnostic assessments as well as a set of training exercises. The assessments provide a diagnostic tool for clinicians, teachers and parents to identify areas of deficit. The exercises provide a training environment through which children learn to improve their face processing skills using a number of engaging games. A single exercise can be used to train a wide range of face processing skills, while each exercises presents training material in a unique way.

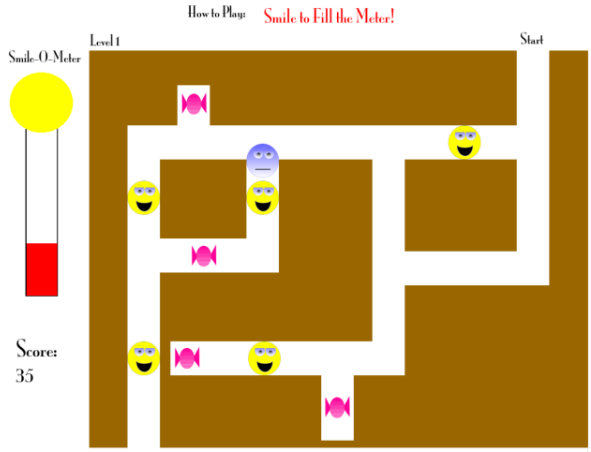
Preliminary findings from a randomized clinical trial indicate that children who played LFI! for 20 hours over a twelve-week intervention period showed reliable, (t (59) = 2.61, p=.006; Cohen's d = .69) gains in their ability to recognize the expression and identity of a face using holistic strategies. These results show that "face expertise", like other forms of perceptual expertise, can be enhanced through direct and systematic instruction. Although these preliminary results are promising, a limitation of the LFI! program is that it only uses static training stimuli and does not incorporate the subjects' own dynamic facial productions. In light of evidence suggesting that individuals with autism have impaired or atypical facial expression production abilities [23] the shortcomings of LFI! could be addressed by incorporating dynamic interactions.

4. Training facial expression expertise

While LFI! provides a comfortable and engaging training environment for children with ASD, it only addresses recognition, not production of expressive faces. Relative to neurotypical individuals, individuals with autism are less likely to spontaneously mimic the facial expressions of others [24] and their voluntary posed expressions are more impoverished than those generated by typically developing individuals [25]. Several studies have already shown that a human interventionist can effectively train individuals with an ASD on facial expressions, including some generalized responding [26], providing even greater impetus for our goal of using software for this training. Moreover, training in facial expression production may improve recognition, as motor production and mirroring may be integral to the development of recognition skills. As such, entangling facial expression perception and production training may prove more fruitful than either training paradigm would alone.

4.1. SmileMaze

We have incorporated the real-time face recognition capabilities of CERT into the LFI! treatment program in a prototype training exercise called SmileMaze. The goal of the exercise is to successfully navigate a maze while collecting as many candies as possible. The player controls a blue pacman-like game piece using the keyboard for navigation (up, down, left, right) and uses facial expressions to move their game piece past obstacles at various points within the maze. As shown in Figure **1**a, the player's path is blocked by a yellow smile gremlin. In order to remove the gremlin and continue along the maze path, the player must produce and maintain a smile for a fixed duration of time.



a.

b.

Figure 1. a. Screenshot of the SmileMaze game b. Sample interaction

Video input is captured using a standard laptop video camera and is continuously analyzed by CERT. If CERT detects a smile, the Smile-O-Meter (Figure **1**a, red bar left side) begins to fill. As long as CERT continues to detect a smiling face, the Smile-O-Meter will continue to fill. However, the moment a non-smile is detected the Smile-O-Meter will cease to fill until a smile is once again detected. Once the Smile-O-Meter has been filled, the obstacle is removed and the player may pass. Feedback regarding the facial expression being detected is provided via the game piece, which will change to show the expression being detected.

Informal field-testing indicates that children with ASD, neurotypical children and adults enjoy playing the SmileMaze exercise. Observations suggest that the game initially elicits voluntary productions of smile behaviors. However, users find the game to be naturally entertaining and amusing thereby evoking spontaneous smiling expressions during game play. SmileMaze demonstrates the connection between voluntary and involuntary expression actions in a gaming format where voluntary productions can lead to involuntary productions and changes in affective state.

4.2. Training in a real-world environment

CERT has been shown to perform with high accuracy in a wide variety of real-world conditions; however, the nature of a training environment implies that the system will need to cope with a substantial degree of atypical expressive input. Indeed, pilot testing demonstrated that players enjoy trying to trick the system, posing with odd expressions. While this makes things more difficult for the system, we wish to encourage this sort of exploratory and entertaining behavior. In order to provide a predicable and intuitive user interaction we have incorporated a number of design techniques into SmileMaze to ensure that CERT can cope with atypical interactions. Not only is a stable and robust system desirable from a training standpoint, it is also of paramount importance for a natural and comfortable user interaction.

CERT was designed to work with full-frontal face images. To account for this, we designed SmileMaze such that players naturally orient themselves facing the camera, ensuring that a full-frontal face image can be captured. This was achieved by using a camera mounted at the top center of the computer screen as opposed to a stand-alone camera beside the monitor. This allows players to interact with the system using the video camera without explicitly directing their behavior toward the camera. Indeed, pilot testing showed that players seldom explicitly look for, or at the camera when producing facial expressions; rather, their focus is directed at the computer monitor. This provides a tight coupling between user input and system feedback, resulting in a natural and intuitive interaction.

As mentioned previously, pilot testing showed that players enjoyed trying to trick the system with unnatural facial expressions. To assist CERT in accurately labeling facial expressions we always provide a target expression for players to produce. This limits the scope of the possible expressions CERT is required to detect at any given point in time. By providing a target expression, CERT is only required to detect a smiling face, while any other facial expression is deemed to be a failure to produce the appropriate expression. A binary decision (is a smile present or not) reduces the decision space, resulting in very robust expression detection.

5. Future work

Here we do not discuss the intended future work in automated face and expression recognition; rather, we would like to focus on some of the behavioral questions that we can begin to explore by leveraging technologies on the cutting edge of automated real-time face recognition.

5.1. Extending the LFI! training program

We developed SmileMaze as a proof of concept prototype that could be used to explore the dynamics of training expressive production alongside perception. We noted that participants were more engaged in their tasks and that they found the training exercises more fun if they could actively produce expressions as opposed to passive viewing. Formal measures of the benefits of including production into the LFI! training program have not yet been collected as we only have a single expression production-based exercise. However, with the addition of

more production-based exercises we intend on quantifying the benefits of a perception/production based training program.

In extending the variety of production-based exercises we are also able to address a number of open scientific questions. One such question relates to stimulus familiarity. Using CERT, we are now able to capture and quantify training stimuli from the participant's environment. Parents, teachers, siblings and friends can all have images captured via web-cam. Captured images can be quantified and labeled by CERT, and integrated into the training stimuli set. While this may provide a more engaging environment for the participant, it may benefit in other ways as well. Participants may be able to bootstrap skills acquired learning from familiar faces onto novel faces, generalizing the expertise they have learned. Also, using familiar faces from the participant's environment may help with the translation from skills learned in training exercises to their application in the real world. Learning facial expressions using images of mom and dad can be directly applied and experience in the home environment.

A second question we can explore in expanding the diversity of production-based exercises relates to the generality of expressions. We are now able to develop an "Emotion Mirror" application (Figure 2) in which players control the expressions of a computer-generated avatar and/or images and short video clips of real faces. This supports a highly important skill, namely expression invariance. Here, participants can explore the same expression on difference faces. This aids in training a generalized understanding of facial expressions. It also knits expressive production and perception as it is the participant's own face that drives the expressions shown on the avatar and/or image.

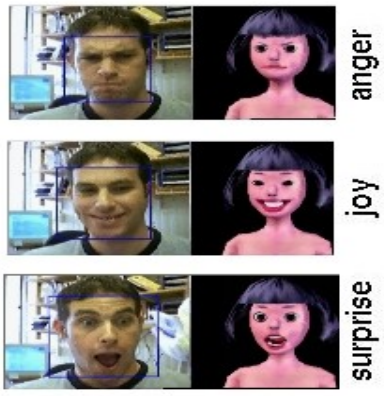


Figure 2. Emotion Mirror: An avatar responds to facial expressions of the subject in real-time.

5.2. An extended expression assessment battery

One of the primary contributions of LFI! was the development of a face perception assessment battery. With the addition of CERT, we are now able to augment the LFI! perception battery with measures of expression production. In preliminary research we tested 15 typically developing children ages 4-10 on a set of facial expression production tasks, including imitation from photograph or video, posed expressions with a vignette (e.g., no one gave Maria presents on her birthday. She is sad. Show me sad), and spontaneous expressions (e.g. children are given a clear locked box with a toy inside and the wrong key), as well as smiles recorded during SmileMaze. Preliminary analysis was performed using CERT to characterize the distribution of facial expression productions of normally developing children. This distribution can then be used to measure the expressive production of individuals in a range of tasks and scenarios.

5.3. Translating trained skills into the world

The main goal of an intervention like LFI! is to train and develop skills that translate into improvements in living standards. The goal of LFI! is not just to have participants show improvements on assessment batteries, but to show improvements in their real-world skills. Recognizing when a friend or business customer is unhappy and understanding what that means is crucial to social interaction.

With the inclusion of expression production into the LFI! training and assessment battery we are now able to probe the integration of expressions and emotions at a level not previously possible. Physiological measures such as heart rate and skin conductance have been shown to be sensitive to affective state [27]. It has also been shown that the production of facial muscle movements associated with an emotion produce automatic nervous system responses associated with those emotions [28].

Given that CERT allows us to train and assess the production of facial expressions, we are now able to measure changes in the physiological affects of expression production as in indicator of the integration of expressions and their meanings. While this may not provide conclusive evidence of development beyond production and perception into a cognitive understanding it would provide a strong contribution towards a convergence of evidence not otherwise possible.

6. Conclusions

The CERT system has been shown to encode face activation units and label basic facial expressions in real-time with a high degree of accuracy under a variety of environmental conditions. Also, the LFI! intervention program has been shown to affectively diagnose face processing deficits and improve those skills through a face training curriculum. We have combined these two technologies into a facial expression production training exercise, SmileMaze. While still in a prototype phase, pilot tests strongly suggest that including expressive production into the LFI! program is a great benefit to its users. Further, and of interest to the scientific community, the inclusion of automatic expression recognition allows a number of high-impact and previously intractable issues to be explored.

Acknowledgements

Support for this work was provided in part by NSF grants SBE-0542013 and CNS-0454233. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the
author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] R. Adolphs, L. Sears, and J. Piven, "Abnormal Processing of Social Information from Faces in Autism," *J. Cogn. Neurosci.*, vol. 13, pp. 232-240, 2001.

[2] P. Ekman and W. Friesen, *The Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press, 1976.

[3] D. Gianluca, B. Marian Stewart, C. H. Joseph, E. Paul, and J. S. Terrence, "Classifying Facial Actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 974-989, 1999.

[4] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, pp. 615-625, 2006.

[5] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, pp. 137-154, 2004.

[6] I. Fasel, B. Fortenberry, and J. Movellan, "A generative framework for real time object detection and classification," *Computer Vision and Image Understanding*, vol. 98, pp. 182-210, 2005.

[7] T. Kanade, J. F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," presented at Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, 2000.

[8] M. S. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, pp. 253-263, 1999.

[9] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," presented at Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, 2005.

[10] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," presented at Systems, Man and Cybernetics, 2004 IEEE International Conference on, 2004.

[11] M. S. Bartlett, G. Littlewort, E. Vural, K. Lee, M. Cetin, A. Ercil, and J. Movellan, "Data mining spontaneous facial behavior with automatic expression coding," *Lecture Notes in Computer Science*, vol. 5042, pp. 121, 2008.

[12] J. Whitehill, M. S. Bartlett, and J. Movellan, "Automated teacher feedback using facial expression recognition," in *CVPR workshop*, 2008.

[13] A. American Psychiatric, D.-I. American Psychiatric Association. Task Force on, and PsychiatryOnline.com, *Diagnostic and statistical manual of mental disorders DSM-IV-TR*. Washington, DC: American Psychiatric Association, 2000.

[14] A. Klin, S. S. Sparrow, A. de Bildt, D. V. Cicchetti, D. J. Cohen, and F. R. Volkmar, "A Normed Study of Face Recognition in Autism and Related Disorders," *Journal of Autism and Developmental Disorders*, vol. 29, pp. 499-508, 1999.

[15] P. Hobson, J. Ouston, and A. Lee, "What's in a face? The case of autism," *British Journal of Psychology*, vol. 79, pp. 441-453, 1988.

[16] R. M. Joseph and H. Tager-Flusberg, "An Investigation of Attention and Affect in Children with Autism and Down Syndrome," *Journal of Autism and Developmental Disorders*, vol. 27, pp. 385-396, 1997.

[17] A. L. Lewy and G. Dawson, "Social stimulation and joint attention in young autistic children," *Journal of Abnormal Child Psychology*, vol. 20, pp. 555-566, 1992.

[18] D. Fein, D. Lueci, M. Braverman, and L. Waterhouse, "Comprehension of Affect in Context in Children with Pervasive Developmental Disorders," *Journal of Child Psychology and Psychiatry*, vol. 33, pp. 1157-1162, 1992.

[19] J. W. Tanaka, S. Lincoln, and L. Hegg, "A framework for the study and treatment of face processing deficits in autism," in *The development of face processing*, H. Leder and G. Swartzer, Eds. Berlin: Hogrefe, 2003, pp. 101–119.

[20] M. Heimann, K. Nelson, T. Tjus, and C. Gillberg, "Increasing reading and communication skills in children with autism through an interactive multimedia computer program," *Journal of Autism and Developmental Disorders*, vol. 25, pp. 459-480, 1995.

[21] M. Moore and S. Calvert, "Brief Report: Vocabulary Acquisition for Children with Autism: Teacher or Computer Instruction," *Journal of Autism and Developmental Disorders*, vol. 30, pp. 359-362, 2000.

[22] M. Ferrari and S. Harris, "The limits and motivating potential of sensory stimuli as reinforcers for autistic children," *Journal of Applied Behavioral Analysis*, vol. 14, pp. 339-343, 1981.

[23] D. N. McIntosh, A. Reichmann-Decker, P. Winkielman, and J. L. Wilbarger, "When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism," *Developmental Science*, vol. 9, pp. 295-302, 2006.

[24] D. N. McIntosh, A. Reichmann-Decker, P. Winkielman, and J. L. Wilbarger, "When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism," *Dev Sci*, vol. 9, pp. 295-302, 2006.

[25] S. J. Rogers, S. L. Hepburn, T. Stackhouse, and E. Wehner, "Imitation performance in toddlers with autism and those with other developmental disorders," *J Child Psychol Psychiatry*, vol. 44, pp. 763-81, 2003.

[26] J. DeQuinzio, D. Townsend, P. Sturmey, and C. Poulson, "Generalized imitation of facial models by children with autism," *Journal of Applied Behavior Analysis*, vol. 40, pp. 755-9, 2007.

[27] J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, *Handbook of psychophysiology*. Cambridge, UK; New York, NY, USA: Cambridge University Press, 2000.

[28] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, pp. 1208-1210, 1983.

Exploring Behavioral Expressions of Player Experience in Digital Games

Wouter van den Hoogen
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven,
The Netherlands
W.M.v.d.Hoogen@tue.nl

Wijnand IJsselsteijn
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven,
The Netherlands
W.A.IJsselsteijn.nl@tue.nl

Yvonne de Kort
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven,
The Netherlands
Y.A.W.d.Kort@tue.nl

Abstract

This paper describes a first exploration of human motor behavior that may be associated with player experiences in digital games. Evidence from literature suggests that patterns in pressure and postural movement data may be indicative for experiences such as interest, arousal, frustration and boredom. In the current study we explore the relation between behavioral measures and people's emotional experience during game play. Results from the study presented in this paper indicate that the intensity of people's actions (e.g. pressure exerted on the mouse) and bodily movement relates to several experiences during game-play, including frustration. However, the results show that these behavioral measures do not exclusively relate to one specific experience. Rather, the results imply these behavioral measures to relate to the level of arousal and level of dominance felt during game-play. From these results it is evident that behavioral measures have a clear application potential. This study presents a starting point in the development of a set of behavior-based measures of player experiences. Establishing sensitivity and validity of such measures can be regarded as the necessary first step in the process of creating an emotionally adaptive game.

1. Introduction

One of the main challenges facing the digital games research community is the development of a coherent and fine-grained set of methods and tools that enable the measurement of entertainment experiences in a sensitive, reliable and valid manner. Measures that capture users' emotions and experiences during gameplay will substantially enhance our understanding of game elements that are particularly engaging and motivating. This will likely aid theory development by allowing a much more direct coupling between specific game design patterns [1] and player experiences. Moreover, understanding gameplay at its base level will allow game designers to introduce those design elements in a game which are known to elicit the most engaging experiences, based on an understanding of what the player will be experiencing at each point in the game. Eventually, the output of continuous measures of player experiences may become real-time input to the game engine, allowing the game's artificial intelligence to adjust to the player's affective or cognitive state at any point during gameplay.

It should be noted that in the large body of literature on media reception and reaction processes, the behavioral impact of media is usually discussed in terms of how media affect behavioral tendencies after episodes of media exposure. For example, a significant body of digital games research is looking at potential associations between exposure to violent games and the development and manifestation of antisocial (e.g., aggressive) behaviors [6]. However, when we refer to behavioral responses in the current paper, we are referring to naturally occurring physical and social behaviors as they are exhibited during an episode of gameplay, as a direct response to unfolding game events and/or social interactions among multiple game participants.

The current paper sets out to describe a first exploration of behavioral expressions that could serve as real-time indicators of experiences related to playing digital games. In this paper, we focus primarily on pressure patterns exerted on a physical control device, and postural responses. Based on this exploration, we present our progress in developing a set of behavior-based measures of such player experiences and their application in an

978-1-4244-2154-1/08/$25.00 ©2008 IE

experimental study.

1.1. Flow, frustration and boredom

Csikszentmihalyi [4,5] studied what makes experiences enjoyable to people. He was interested in people's inner states while pursuing activities that are difficult, yet appear to be intrinsically motivating, that is, contain rewards in themselves – chess, rock climbing, dance, sports. In later studies, he investigated ordinary people in their everyday lives, asking them to describe their experiences when they were living life at its fullest, and were engaged in pleasurable activities. He discovered that central to all these experiences was a psychological state he called flow, an optimal state of enjoyment where people are completely absorbed in the activity. Flow is a state where someone's skills are well balanced with the challenges posed by a task. It is characterized by a deep concentration on the task at hand, a perceived sense of control over actions, a loss of preoccupation with self, and transformation of one's sense of time.

Flow certainly sounds familiar to frequent players of computer games. Digital games provide players with an activity that is goal-directed, challenging and requiring skill. Most games offer immediate feedback on distance and progress towards the (sub)goals, through, for instance, score keeping, status information (e.g., a health indicator), or direct in-game feedback. When a game is effective, the player's mind can enter an almost trance-like state in which the player is completely focused on playing the game, and everything else seems to fade away - a loss of awareness of one's self, one's surroundings, and time. It is the experience that is strongly connected to what gamers and game reviewers commonly refer to as the 'gameplay' of a game, i.e., the somewhat ambiguous term describing a holistic gaming experience, based on a fluent interaction with all active gaming elements, the progression of challenges offered, and the ability of a game to continuously command the attention of a player.

Sweetser and Wyeth [19] have adopted and extended Csikszentmihalyi's conceptualization of flow in their 'GameFlow' model of player enjoyment, formulating a set of useful design criteria for achieving enjoyment in electronic games – see also [8]. Csikszentmihalyi's original work on flow suggests that these peak experiences are quite rare – the exception rather than the rule. Nevertheless, the flow model of game enjoyment clearly illustrates the importance of providing an appropriate match between the challenges posed and the player's skill level. The flow experience can easily break down when the player's skills systematically outpace the challenges the game can offer (leading to boredom) or when game challenges become overwhelming in light of the available skills (resulting in frustration). Challenge is probably one of the most important aspects of good game design, and adjusting the challenge level to accommodate the broadest possible audience in terms of player motivation, experience and skill is a major challenge for current game designers.

Being able to detect frustration and boredom is of importance as indicators of when a person is not experiencing flow, but also, and perhaps more interestingly, because successful games strike a balance between positive and negative emotions (see, e.g., [16]). This is in line with the view that games are often being designed with the aim to develop a negative emotion in the face of challenge, only to be followed by a positive emotional peak when the challenge is overcome [9]. In sum, behavioral indicators of involvement or interest are required, as well as indicators of both boredom and frustration.

1.2. Behavioral expression of player experiences

Behavioral expressions of subjective states are well known to both lay-people and scientists alike. A host of observable and expressive physical behaviors are associated with emotional states. We tend to smile at something funny, move towards something or somebody we like, jump up when startled, hide our heads when scared, or make strong gestures when frustrated. There are a number of behavioral responses where the human motor system may potentially act as a carrier for the player experiences discussed previously.

Mota & Picard [14] demonstrated that postural patterns can be indicative of learner interest. They developed a system to recognize postural patterns and associated affective states in real time, in an unobtrusive way, from a set of pressure sensors on a chair. Their system is reportedly able to detect, with an average accuracy of 87.6%, when a child is interested, or is starting to take frequent breaks and looking bored. Thus, the dynamics of postures can distinguish with significant reliability between affective states of high interest, low interest and boredom, all of which are of relevance to a gaming situation as well.

Clynes [2,3] investigated the patterns of motor output of people asked to deliberately express certain emotions through the motor channel (usually a finger pressing on a measuring surface he dubbed the 'sentograph'). He found that there are distinguishable, stable patterns of pressure and deflection for emotions such as anger, hate, grief, love, and joy, transcending barriers of culture and language [2]. Support for Clynes' original findings has been varied. Trussoni, O'Malley and Barton [21] failed to replicate Clynes' findings using an improved version of the sentograph. Although they did find distinguishable patterns associated with certain emotions, a significant

correlation with Clynes' original sentograms [2] was absent, throwing doubt on the universality of sentic patterns. Hama and Tsuda [7], on the other hand, did find support for the characteristic waveform patterns associated with 'sadness' (long duration of pressure) and 'anger' (strong intensity of pressure). Moreover, in their first experiment, Hama and Tsuda did not inform participants that they were interested in measuring emotions, which raises the interesting possibility that identifiable pressure patterns may be associated with spontaneously generated motor expression of emotions. In particular, the sentic expression of anger is of interest as a potential indicator of gamer frustration.

Research by Mentis and Gay [13] and Park, Zhu, McLaughlin & Jin [15] provide evidence that the force people apply to interface devices can be interpreted as an indicator of negative arousal. Mentis and Gay [13] asked a small number of participants to complete several tasks on a word processor. Later, participants were asked to indicate whether and when they experienced a frustrating event. Their results suggest that higher pressure on the touchpad is associated with a frustrating event. Building on these findings, Park et al. [15] manipulated frustration by asking participants to complete an impossible LEGO assembly task. The instructions for the task and optional online help were presented on a laptop computer, where the pressure exerted on the touchpad was measured. Results indicated that more pressure was exerted on the interface device when participants were encountering problems. Additionally, pressure patterns also correlated with facial expressions showing negative affect, thereby providing evidence that the pressure exerted was indeed related to frustration rather than mere arousal.

Focusing on digital games, Sykes and Brown [20] have investigated the mean pressure exerted by players on a gamepad's button as the difficulty level of a game (Space Invaders) was increased from easy to medium to hard. Their results show that buttons on the gamepad were pressed significantly harder in the hard condition than in either the easy or the medium condition. Although the increase in pressure on the gamepad can be assumed to be associated with higher arousal, Sykes and Brown did not determine whether this arousal was positively or negatively valenced, while both states could plausibly occur in a digital game setting. Notwithstanding this limitation, Sykes and Brown [20] successfully demonstrated that a fairly straightforward behavioral measure such as hand or finger pressure exerted on a button can already be informative about the level of user arousal in gaming situations. In addition, given its relative simplicity, this measure has the potential to be analyzed in real-time and be used to adaptively influence the game dynamics.

2. Linking behavior to player experience

From the literature it is evident that behavioral patterns are likely to be informative for the real-time measurement of player experiences. From a methodological point of view, there are several advantages associated with employing behavioral measures as an indicator of player experiences. First, they are relatively free from subjective bias, because they are generally not under users' conscious control, nor do they require specific instructions from an experimenter (e.g., "please hit the button harder as you get more frustrated") – they occur spontaneously. Secondly, when measured in an unobtrusive fashion, they do not disrupt the player experience. Third, they are time-continuous measures, that is, they are collected as the experience is unfolding, and are as such not reliant on memory or introspection on the part of the participant (unlike self-report measures). Finally, a number of these measures, such as a pressure-sensitive gamepad, could realistically be integrated with existing game technologies. This is a clear advantage when these measures are to be integrated in commercial games, where specialist peripheral hardware will only scarcely be adopted.

In the current study, we want to explore a number of behavioral measures in relation to player experiences. The aim of such an exploration is twofold. First, we need to establish which behavioral measures are sensitive to variations in game dynamics. Second, we need to find out in what way behavioral measures are correlated to player experiences, thereby establishing a potential connection between objective measurements and subjective experience. Behavioral indicators that are demonstrated to be both sensitive to experimental manipulations and sensibly related to player experiences can subsequently be deployed in closing the loop between the player and the game. That is, successful behavioral indicators of player experience can be used as real-time input data to the game engine, dynamically adapting the game to the player's experiential state. The current study should thus be regarded as the necessary first step in the process of creating an emotionally adaptive game, establishing sensitivity and validity of behavioral indicators of player experiences.

In an attempt to link measurable behavior such as postural movements and pressure patterns to people's emotional states during digital game play, we have recently developed several real-time behavior measurement systems, including a pressure-sensitive chair, inspired on the work of Mota and Picard [14], and a pressure sensitive mouse and keyboard. Although we have reviewed and tried various off-the-shelf solutions, including VR pressure-sensitive gloves, our fairly straightforward, customized measures allow for more sensitive measurement of various bodily responses, are not overly obtrusive, and can be easily integrated with

existing gaming devices. Moreover, the combination of *multiple* behavioral indicators can reduce uncertainty or ambiguity associated with a single indicator, resulting in increased robustness and wider applicability of the total set of measures. Limitations particular to one measure may be overcome or compensated by using corroborating evidence emerging from another measure.

In the study reported in this paper we therefore decided to use multiple behavior measurement systems in conjunction with self report measures of people's game-play. Within this study we have used customized levels of a digital game (Half Life 2) to induce boredom, enjoyment, and frustration. By inducing these player experiences, we do not need to infer these states, nor wait for their spontaneous occurrence. Moreover, such a manipulation is expected to result in much needed variation in types of experiences. This will allow us to more reliably associate behavioral response patterns with affective states (see also [17]).

3. Method

The experiment was conducted in the Game Experience Lab at Eindhoven University of Technology. The first person shooter game Half Life 2 was modified such that game difficulty was either easy, moderate, or hard, according to a within groups design. After each level participants filled in a questionnaire including the self report measures aimed to measure player experience. The game was played on a Dell XPS PC equipped to cope with the demands of the game and was connected to a 20" TFT-screen.

3.1. Participants

Thirty-two participants (five females) aged between 17 and 46 (Mage = 22.42 years, SD = 5.57 years) took part in the experiment. All participants at least occasionally played first person shooter (FPS) games, but a substantial part consisted of more frequent players. Participants received 10€ for their time.

3.2. Procedure

Upon entering the lab, participants were welcomed by the experiment leaders. The experiment leaders gave a brief overview of the progression of the experiment. Participants signed the consent form (allowing video observations and psychophysiological measures to be taken), were seated at a desk where the game-PC was installed, and were connected to psychophysiological sensors and an accelerometer. After reading brief instructions related to the use of the controls in the game, participants played the three customized levels of the FPS game Half Life 2. After each level, participants rated their experiences during game-play on a range of self-report measures administered on a separate laptop PC. The order in which the levels were played was counterbalanced. Participants were given ten minutes to play each of the levels with the exception of the easy level. Because more experienced players usually finished the easy level in less than ten minutes, we fixed the playtime for this level at eight minutes. At the end of the session, participants were paid, debriefed, and thanked for their participation.

3.3. Measures

The study was designed to relate behavioral responses to self reported experiences during game play. Consequently, during the study we measured both people's self reported experience of each level played and measured their behavior using a range of behavioral measurement tools. The measures are described in more detail below.

3.3.1 Self report measures

Self report measures used in the study included the Self Assessment Manikin (SAM) scale [10;18] and the in-game version of the Game Experience Questionnaire [22] recently developed by the Game Experience Lab at Eindhoven University of Technology. Further, we included a manipulation check for the level of difficulty.

3.3.1.1 SAM-scale

The SAM scale is a visual self report scale developed by Lang [10] and based on Mehrabian and Russell's [12] Pleasure-Arousal-Dominance (PAD) theory. The SAM-scale visualizes the three PAD-dimensions. Each dimension is depicted through a set of five graphic figures (manikins) and for every dimension respondents have to indicate which figure corresponds best with their feelings on a nine point scale. The first dimension P (displeasure/pleasure) ranges from extreme sadness to extreme happiness. The second dimension A (non-arousal/arousal) ranges from very calm or bored to extremely stimulated. The third dimension D (submissiveness/dominance) ranges from a feeling of being controlled or dominated to a feeling of total control. Additionally, we included a SAM-based measure of presence developed by Schneider et al. [18] as a fourth emotion dimension that possibly applies to digital game experience. This dimension ranges from a feeling of total presence to a feeling of total absence. For each SAM dimension we asked participants to indicate, on a 9-point scale listed below the graphical presentation, which manikin corresponded with their experiences during game-play. Scale values ranged from -4 to 4, with ascending scores corresponding to higher pleasure, higher arousal, higher dominance and lower presence ratings.

3.3.1.2 *In-Game GEQ (iGEQ)*

After each level we administered the in-game Game Experience Questionnaire (iGEQ) consisting of seven dimensions with two items per dimension. These dimensions were: Positive affect *(I felt content, I felt good*), Boredom (*I found it tiresome, I felt bored*), Frustration (*I felt irritable, I felt frustrated*), Flow (*I felt completely absorbed, I forgot everything around me*), Challenge (*I felt stimulated, I felt challenged*), Immersion (*I was interested in the game's story, I found it impressive*), and Competence (*I felt skilful, I felt successful*). All GEQ items are measured by means of five point intensity scales with points anchored at not at all (0), slightly (1), moderately (2), fairly (3), extremely (4). For our analyses, we used the mean value of the two items per dimension. We used the iGEQ, the shorter in-game version of the GEQ, because we did not want to interrupt participants too long between the different levels of game-play.

3.3.1.3 *Manipulation check*

The manipulation check included one five point bi-polar statement stating "How easy or difficult did you find it to play the level?" ranging from -2 (too easy to play) via 0 (optimal to play) to 2 (too difficult to play).

3.3.2 *Behavioral measures*

During the game-play we measured people's movement on the chair they were sitting on, measured the movement of their upper body by means of an accelerometer, and measured the force they applied to the mouse. Each of these measures is shortly explained below.

3.3.2.1 *Accelerometer*

For each participant, an accelerometer was attached to the back, at the base of the neck, to automatically capture movement of the upper body. The accelerometer used was a Phidgets 3 axis version measuring tilt on the x, y, and z-axes, and acceleration to a maximum of 3Gs, which is more than enough for the expected movement of the participants during game play. For the analyses we used the accelerometer data converged over all axes (square root of the sum of the squared values for each of the three other axes). Subtracting the mean value across all levels from the individual data values and calculating the absolute value resulted in a metric representing the acceleration as a function of movement in any direction. In addition to the maximum value per level, these values were averaged per level providing an indication of the average movement during each level.

3.3.2.2 *Pressure sensitive chair*

A second automatic indicator of movement was recorded via a pressure sensitive chair. Sitting position and the number of shifts in position are potential indicators of boredom and of interest. In addition to observed and coded sitting position (forward-backward movement) using the video streams, we also employed a custom-built posture-sensitive chair using force-sensitive sensors built into the legs of the chair. This allowed real-time measurement of the forward-backward and sideways movements of the participant during game-play. The sensors used were TekScan pressure sensitive sensors designed to measure up to 25Lbs (approx. 11.3 Kg) of force applied to them (for an image of the chair and the measuring system see Figure 1).

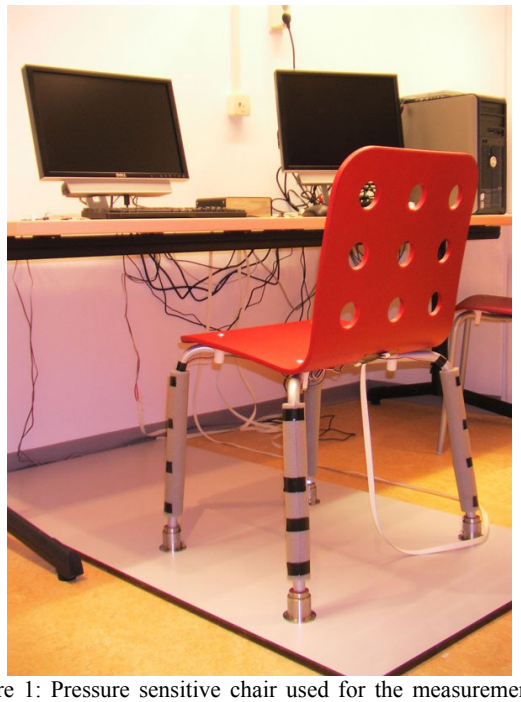


Figure 1: Pressure sensitive chair used for the measurement of people's changes in sitting position.

For the purposes of the current study we calculated the maximum range of forward-backward movement on the chair by subtracting the minimum value (most backward position) from the maximum value (most forward position). This measure was calculated for each of the levels played. As there are likely large individual differences in the rate of movement we applied a range correction to the measures. That is, the values of the range of movement for each level were divided by the maximum range across all levels for that individual. This procedure was used as this is advised for the use of galvanic skin response (GSR) data [11] which has similar properties and dependencies on individual differences to our automatically captured behavioral measures. Additionally, it neutralizes potential differences in sensitivity of the

pressure sensors (e.g. due to differences in weight of the participants), and allows comparison across individuals.

3.3.2.3 Pressure sensitive mouse

The mouse was equipped with two Flexiforce sensor designed to measure up to 1Lbs [approx. 453.6 grams] of force applied to them, mounted on top of two buttons. To increase the likelihood that the participants would press on the sensors when operating the mouse the paddles were reduced in size and the sensors were topped with a small rubber patch. This patch raised the surface of the sensor over the rest of the paddle and discriminated the surface texture of the paddles. The patch thus naturally invited people to keep their fingers on top of the sensors (see Figure 2 for a view of the augmented mouse).

Figure 2: Pressure sensitive mouse measuring force applied to the mouse during game-play.

The mouse pressure data were recorded continuously allowing for synchronization of the force on the input devices with discrete in-game events. The data can be aggregated over lengths of time, e.g., complete sessions. This provides opportunities for event-based analyses, and correlation analyses with self-report measures. For analyses of the force applied to the mouse, two measures were constructed. The first measure was constructed using the maximum value of force applied to the left mouse button per level. As with the chair we applied a range correction to the values. That is, the maximum value of each level was divided by the overall maximum value of force during game-play. Again this was done to reduce individual differences in the force people apply on an interface device, allowing comparison between participants. The second measure constructed was the average force applied to the mouse based on the maximum force per event, thus excluding all values between the onset and end of the mouse press other than the maximum force. Again like the maximum mouse force, the mean force was based on the range corrected values.

4. Results

First the results of the manipulation check will be presented, followed by the results of the behavioral measures. Although the results of the self-report measures will be reported elsewhere, the correlation between the self-report measures and the behavioral measures are reported in this paper, since this provides an indication of the validity of the behavioral measures.

4.1. Manipulation Check

The three levels used in this study were designed to represent an easy, a challenging, and a hard level in terms of difficulty, ideally inducing boredom, flow/ enjoyment, and frustration. In order to establish the effect of the manipulation we conducted a repeated measures ANOVA on the one-item manipulation check '*How easy or difficult did you find it to play the level?*'. This analysis showed significant differences between each of the three difficulty levels in the expected directions ($F(2,30)=120.77$, $p<.001$). The easy level was rated as the "easiest" (M= -1.47, SD= 0.84) followed by the moderate level (M= -0.5, SD= 0.95), and the hard level (M=1.09, SD= 0.73) as the most difficult level to play. This result thus provides initial confirmation that the difficulty manipulation was effective.

4.2. Behavioral measures

4.2.1 Accelerometer

Using a repeated measures ANOVA we analyzed the effects of the level of difficulty on both the maximum and mean scores acquired. A Greenhouse Geisser correction for the repeated measures ANOVA was used, correcting for violations of sphericity. The results indicate that there is no difference in the maximum accelerometer data between the levels (see Table 1). The mean accelerometer value did, however, differ between the levels ($F(1.61,29.39)=10.69$, $p<.001$). The mean value for the hard level proved to be highest and significantly different from both the moderate and the easy level. This implies that in the hard level participants, on average, moved more strongly than in the other levels.

4.2.2 Sitting position

The second behavioral indicator of player movement was acquired via the sensors in the pressure sensitive chair. This indicator takes into account not only the movement of the upper part of the body, but rather the center of gravity of the body as a whole. Sitting position was analyzed using the corrected range from the forward-backward position on the chair. The results from the

repeated measures ANOVA showed the range in movement on the chair to differ significantly between the levels ($F(2,29)=5.52$, $p=.006$) with the hard level having the highest score and differing from the easy and moderately difficult levels (see Table 1). In line with the accelerometer results, this implies that movement was strongest in the most difficult level.

4.2.3 Mouse pressure

Sensors on the left mouse buttons measured the force with which players made each mouse click. The force applied to the mouse was analyzed using both the maximum mouse force and the mean mouse peak force. Both indicators increased with the difficulty of the game level (see Table 1). The maximum mouse force differed significantly between the levels ($F(2,29)=11.72$, $p<.001$), with the hard level differing significantly from both the easy and moderately difficult levels. As for the maximum mouse force, mean mouse force was highest in the hard level. The effect was however only marginally significant, with the easy and hard level differing from each other. The results show that, on average, people applied most force on the mouse in the difficult level. Similar to the accelerometer and chair results, this result implies that the behavior was again most intense in the most difficult level.

TABLE 1: MEANS OF BEHAVIORAL MEASURES PER LEVEL OF DIFFICULTY WITH STANDARD DEVIATIONS IN PARENTHESES
(† = MARGINALLY SIGNIFICANT, *P<.05, ** P<.01, *** P<.001)
([a,b,c]: DIFFERS SIGNIFICANTLY (PAIRWISE COMPARISONS, P < .05) FROM EASY LEVEL[a], MODERATE LEVEL[b], HARD LEVEL[c])

	Easy	Moderate	Hard
Mean Accelerometer ***	0.0036 [c] (.0008)	0.0037 [c] (.0008)	0.0041 [a,b] (0.0012)
Maximum Accelerometer	0.13 (0.18)	0.13 (0.09)	0.13 (0.07)
Chair Movement **	0.59 [c] (0.31)	0.56 [c] (0.32)	0.81 [a,b] (0.28)
Mean Mouse Force †	0.11 [c] (0.059)	0.12 (0.067)	0.13 [a] (0.069)
Maximum Mouse Force ***	0.45 [c] (0.31)	0.54 [c] (0.32)	0.84 [a,b] (0.28)

4.3. Correlations between self report measures and behavioral measures

The results demonstrate that the behavioral measures used in this study (mouse force, movement on a chair, and upper body movement) related to the level of difficulty. More importantly, with the exception of the maximum accelerometer value they were highest when the level of difficulty was highest, in line with previously reported findings (e.g. [20]). However, as earlier research makes clear, exactly what these measures (e.g. mouse force) indicate is unclear. Although mouse force has previously been associated with frustration [13], in gaming this measure may signify both pleasurable challenge as well as frustration. In order to connect the behavioral measures to player experience, we included the iGEQ, and the SAM. In this section we present the correlations between the behavioral measures found to be sensitive to the manipulation of difficulty and the self report measures.

For thoroughly exploring correlations between variables, sufficient variation is needed. Since the experimental levels were explicitly created to induce a specific experience, variance within each level was only modest. For this reason we restructured the data such that the different experimental levels were treated as separate cases, creating three rows of data for each participant. By exploring correlations across levels we created variation in the different measures enabling us to report reliable conclusions about how the self report measures are related to the behavioral measures.

TABLE 2: CORRELATIONS BETWEEN BEHAVIORAL MEASURES AND IGEQ.
(† = MARGINALLY SIGNIFICANT, * P<.05, ** P<.01, ***P<.001)

	Maximum Mouse Force	Mean Accelerometer	Chair Movement
Immersion	.063	-.110	.051
Competence	-.392 ***	-.197 †	-.289 **
Neg. Affect (boredom)	-.262 *	-.052	-.226 *
Flow	.205 *	-.048	.199 †
Frustration	.335 ***	.148	.356 ***
Challenge	.399 ***	.302 **	.252 *
Pos. Affect	-.141	.057	-.240 *

From Table 2 it is evident that the behavioral measures are correlated with multiple dimensions of the iGEQ, rather than only one as previous research suggests [13;15]. The results thus imply the behavioral measure to be related to more than only one specific emotion (such as frustration). Most notably, there is a large overlap in direction and magnitude of the correlations with the iGEQ dimensions of both the Maximum Mouse Force and Chair Movement.

Additionally, the behavioral measures are negatively correlated with items that can be interpreted as being low arousal experiences (Positive Affect, Boredom, and Competence), while they are positively correlated with items signaling higher arousal states (Frustration, Challenge, and to a lesser extent Flow). This suggests that the intensity of behavior (chair movement and pressure on the interface device) is an indicator of arousal as underlying physiological state of the person playing the

game. Correlations of the behavioral measures with the SAM are indeed consistent with this interpretation. As can be seen in Table 3 the behavioral measures are positively correlated with the arousal dimension and negatively correlated with the dominance dimension. Indeed in the context of game-play it seems that these two are to some extent each other counterparts. As people feel they are more dominated by the game (i.e. lose control) they will likely get more aroused through this challenge.

In sum, the correlations of the behavioral measures with the SAM and iGEQ dimensions show that the behavioral measures are likely indicators of arousal, more so then they can be interpreted as indicators of one specific emotion or experience.

TABLE 3: CORRELATIONS BETWEEN BEHAVIORAL MEASURES AND SAM.DIMENSIONS
(† = MARGINALLY SIGNIFICANT, * P<.05, ** P<.01, ***P<.001)

	Maximum Mouse Force	Mean Accelerometer	Chair Movement
Pleasure	-.150	.073	-.160
Arousal	.189 †	.219 *	.222 *
Dominance	-.301 **	-.236 *	-.273 **
Presence	.053	.061	-.038

5. Conclusions

In the study presented in this paper we investigated the potential for multiple behavioral measures as indicators of people's game-play experience. Using an experiment in which difficulty level was manipulated, we have found automatically captured body movement and pressure on the interface device to be highest in the most difficult level. These findings are in line with our hypotheses, and support and extend earlier findings by Sykes and Brown [20] who found that more pressure was exerted on a gamepad's button as the difficulty level increased. However, Sykes and Brown did not take any self-report measures to help interpret pressure as a measure of player experience. Our results suggest there is no easy one-to-one relation with frustration or enjoyment. Rather, we find that an increase in arousal, be it through increased frustration or challenge, results in a higher level of pressure exerted. A similar pattern emerges for the measures of postural movement.

Combined, our findings suggest that measures of movement and pressure mainly serve as indicators of people's level of arousal. The intensity of action has been found to relate to arousal states: i.e. they were highest when the level was most difficult, correlated positively with the high arousal experiences, were negatively correlated with low arousal experiences, and were positively correlated with SAM arousal and dominance scales. Our findings do not support suggestions made in previous research that pressure exerted on a mouse (or touchpad) is *exclusively* associated with the experience of frustration [13, 15]. It is important to note that such results have been obtained in productivity oriented tasks (word processing task, LEGO assembly task), where frustrating events are explicitly included (e.g., a task that is impossible to complete), and positive challenge is lacking. In contrast, a digital game such as Half Life 2 allows for a more varied spectrum of challenges, some of which add to the excitement of the game, others leading to frustration. Thus, more force applied to the interface device cannot be simply translated to higher levels of frustration. The study presented in this paper is a first exploring the relation between multiple behavioral measures and multiple self reported dimensions of game-play. Importantly, our research show not *only* mouse force to correlate with multiple experiences. These relations are evident for chair movement, and to a lesser extent upper body movement as well. Further, the relations appear to be consistent across the behavioral measures included in our experiment, in all cases the intensity of the behavior appears to relate to the level of arousal that players experience. The consistency of these findings bode well for behavioral indicators as potential input data to game engines.

Having made a first step in determining the sensitivity and validity of a number of behavioral measures, we will next turn our attention to determining whether these relations are stable over time (test-retest reliability), as well generalisable across different games and gaming genres. Moreover, although we have used time-continuous measures, we have analysed them in aggregate form (i.e., means across levels and players). In order to firmly establish whether such measures can be useful as input to emotionally adaptive games, we need to establish sensitivity of the measures at an individual level, and across much shorter time-spans (in the order of seconds rather than minutes). Further, analyses of our current rich dataset, as well as new experiments, are expected to throw light on this issue in the near future.

Acknowledgments

Thanks to Martin Boschman for technical assistance and co-design of the measurement systems. Funding from the EC through the IST Framework 6 FUGA (http://project.hkkk.fi/fuga/) and Games@Large projects (http://www.gamesatlarge.eu/) is gratefully acknowledged.

References

[1] Björk, S. and Holopainen, J. (2004). Patterns In Game Design. Charles River Media.

[2] Clynes, J.M. (1973). Sentics: Biocybernetics of emotional communication. Annals of the New York Academy of Science, 220, 55-131.
[3] Clynes, J.M. (1977). Sentics: The Touch of the Emotions. Anchor Press/Doubleday, 1977.
[4] Csikszentmihalyi, M. (1975). Beyond Boredom and Anxiety. San Fransisco: Jossey-Bass.
[5] Csikszentmihalyi, M. (1990). Flow. The Psychology of Optimal Experience. New York: Harper & Row.
[6] Gunter, B. (2005). Psychological effects of video games. In: J. Raessens & J. Goldstein (eds.), Handbook of Computer Game Studies. Cambridge, MA: MIT Press, pp. 145-160.
[7] Hama, H. and Tsuda, K. (1990). Finger-pressure waveforms measured on Clynes' Sentograph distinguish among emotions. Perceptual and Motor Skills, 70, 371-376.
[8] Holt, R. and Mitterer, J. (2000). Examining video game immersion as a flow state. Paper presented at the 108th Annual Psychological Association, Washington, DC.
[9] Keeker, K., Pagulayan, R., Sykes, J., and Lazzaro, N. (2004). The untapped world of video games. ACM CHI 2004, 1610-1611.
[10] Lang, P.J. (1980). Behavioral Treatment and Bio-Behavioral Assessment: Computer Applications. In Technology in mental health care delivery systems, Sidowski, Joseph B., James H. Johnson, and Thomas A. Williams (Eds.). (119-37). Norwood, NJ: Ablex Publishing.
[11] Lykken, D.T., & Venables, P.H. (1971) Direct measurement of skin conductance: A proposal for standardization. Psychophysiology, 8 (5), 656-672.
[12] Mehrabian, A., & Russell, J.A. (1974). An Approach to Environmental Psychology. Cambridge, MA: The MIT Press.
[13] Mentis, H.M. and Gay, G.K. (2002). Using touchpad pressure to detect negative affect. Proceedings of Fourth IEEE International Conference on Multimodal Interfaces 2002, 406 - 410
[14] Mota, S. and Picard, R.W. (2003). Automated Posture Analysis for Detecting Learner's Interest Level. Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction (CVPR HCI). Available: http://affect.media.mit.edu/pdfs/03.mota-picard.pdf
[15] Park, N., Zhu, W., Jung, Y., McLaughlin, M., and Jin, S., (2005). Utility of haptic data in recognition of user state. Proceedings of HCI International 11. Lawrence Erlbaum Associates. Available: http://imsc.usc.edu/haptics/paper/manuscript_hcii2005_final.pdf
[16] Ravaja, N., Salminen, M., Holopainen, J., Saari, T., Laarni, J. and Järvinen, A. (2004) Emotional response patterns and sense of presence during video games: potential criterion variables for game design. Proceedings of the Third Nordic Conference on Human-Computer Interaction, 339-347.
[17] Scheirer, J., Fernandez, R., Klein, J., and Picard, R.W. (2002). Frustrating the user on purpose: A step toward building an affective computer. Interacting with Computers, 14, 2, 93-118.
[18] Schneider, E.F., Lang, A., Shin, M., & Bradley, S.D. (2004). Death with a story: How story impacts emotional, motivational, and physiological responses to first-person shooter video games. Human Communication Research, 30 (1), 361-375.
[19] Sweetser, P and Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. ACM Computers in Entertainment, 3 (3), 1-24.
[20] Sykes, J. and Brown, S. (2003). Affective gaming. Measuring emotion through the gamepad. ACM CHI 2003, 732-733.
[21] Trussoni, S.J., O'Malley, A., and Barton, A. (1988). Human emotion communication by touch: A modified replication of an experiment by Manfred Clynes. Perceptual and Motor Skills, 66, 419-424.
[22] IJsselsteijn, W.A., de Kort, Y.A.W. & Poels, K. (in preparation). The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games. Manuscript in preparation.

A System to Reuse Facial Rigs and Animations

Verónica Costa Orvalho
Instituto de Telecomunicações
Faculdade de Ciências da Universidade do Porto
Porto, Portugal
veronica@faceinmotion.com
http://www.faceinmotion.com

Abstract

Facial animation in films and videogames are strongly dependent on a fixed rig that is custom created for each character. The rig is defined in the early stages of the development and is conditioned by the characters morphology. We present a portable character rigging system that integrates into current animation production pipelines enabling digital artists to create more lifelike characters in less time about 90-99% faster, when compared to traditional animation techniques. It automatically transfers the rig and animations created in one character to different characters, independent of their shape and appearance. Artists are not forced to use predefined rigs and can preserve the original mesh they created. As a result, the system improves the workflow in CG productions, the modeling and animation teams can now work in parallel.

1. Introduction

Facial animation presents many difficulties (time, cost and complexity constraints) that limit its adoption and usefulness in different situations. *Pighin et al.* [20] discuss the research efforts and main challenges faced by some blockbuster films, and emphasize that facial puppetering and the use of nonlinear rigs are still unexplored issues. Generating realistic face movements is hard, because even with current 3D software, animators cannot capture and control every detail of the face. To obtain the desired realism, traditional animation pipelines have each character separately rigged by hand, a very labor-intensive and time-consuming task. A rig is a set of controls that allows an artist manipulate a character. The character rigging process is analogous to setting up the strings that control a puppet, which in the hands of an experienced digital artist comes to life [21]. Finding a technique that provides accurate and fast rigging remains a challenge.

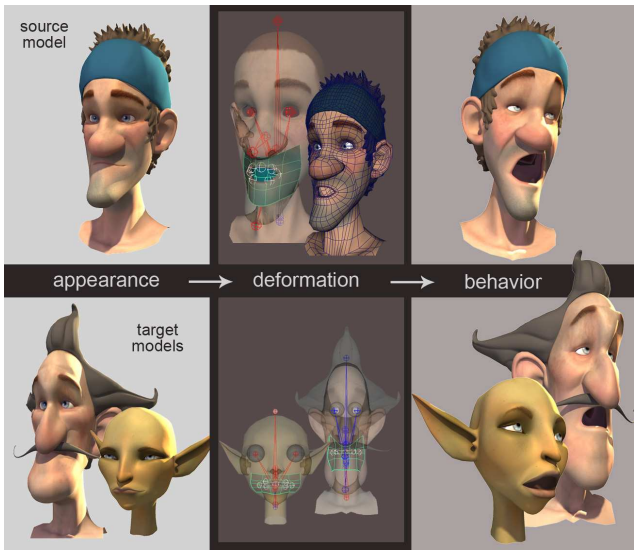


Figura 1. Overview: define the source and target model; adapt the source model geometry to fit the target; transfer attributes and shapes; bind the influence objects and skeleton to the target. The result is a model ready to be animated. (Copyright 2005 Dygrafilms)

Our solution overcomes this problem because it adapts the inner structure of the characters to the shape and facial features of the models. Artist can always keep the mesh they created, so the visual look of the characters is never affected. This is fundamental to guarantee the high quality results required by the entertainment industry (e.g. Playable Universal Capture *Borshukov et al.* [4]), which allows creating and reproducing animations that respect the style and expressiveness of each character, making them unique. Previous methods [18, 7, 27] do not deal directly with the artists needs, and most are oriented towards human look. Our approach is general, so artists can define their own rig and then quickly apply it to different models, even with dis-

parate proportions and appearance (human, cartoon or fantasy). This gives artists complete freedom to manipulate the characters: they can create new animations and not be limited by pregenerated ones. The system we present can easily be integrated into current production pipelines as it is embedded in Maya (Autodesk 2008 [1]). The technology behind the system is described in Orvalho PhD Thesis [19]. It also includes an extensive state of the art analysis related to MPEG-4, FACS another animation techniques. Figure 1 shows an overview of the system

2. Related Work

Facial animation retargetting between dissimilar meshes is not a new problem, but facial rigging retargetting still unexplored. Most work deals only with the geometry of the face and forgets that the key elements to animate a character is the structure underneath the 3D model mesh. Our work differs from previous approaches that focused on transferring animations, because we aim to transfer the complete facial rig, in addition to animations. We also want to allow the reuse of a facial rig in different face models, regardless of the type of the rig. Thus, most existing work is not efficient for real-time animation, are to complex to setup, force the artist to use a fix pre-defined rigs and there is no previous work capable of dealing with our variety of morphologies [24].

Many efforts have been done to retarget or automatically create the body rig of characters [2, 11] and achieved good results, but none have focus on facial rig retargetting. Most facial animation is related to *physically-based*, *geometric deformation* and *performance-driven* methods.

Physically-based methods K. Kahler et al. [14] simulate the contraction and relaxation of human muscles to animate faces. *Yuencheng et al.* [15] used a multiple-layer dynamic skin and muscle model, together with a spring system, to deform the face surface. But these techniques make it hard to define accurate muscle parameters, due to the complexity of human muscles. So, *Sifakis et al.* [24] used non-linear finite element implementation to determine accurate muscle action, captured from motion of sparse facial markers. The method shows the success of performance-driven animation, but it is not clear if it can handle anatomically inaccurate models.

Geometric deformation methods use a variety of techniques to animate faces. Following *Sederberg and Parry* [23], *Chadwick et al.* [5] used Free-Form Deformation (FFD) for layered construction of flexible animated characters, which doesn't require setting the corresponding features on the geometries. *Turner and Thalmann* [26] used an elastic skin model for character animation. Other approaches were introduced for high level geometric control [16, 13] and deformation over 3D model, to help simulate wrinkles [12, 25]. These deformation methods provide artists with easy controls to generate animations, but automating these procedures still needs considerable effort.

Performance-driven methods [4] capture the facial performance of an actor, which can be re-targeted to different face models [9, 6] or blendshapes [7]. These techniques can generate realistic facial motion, but are expensive to use. Also, they are more suited for human beings than imaginary or fantastic characters.

3. Facial Rigging Challenges

"Rigging is the process of taking a static, inanimate computer model and transforming it into a character that an animator can edit frame-by-frame to create motion" [8]. The result is a rig that can be manipulated by a set of controls like a virtual puppet [22] or by motion capture data. Creating the character rig is a very complex, time consuming and labor intensive task. Still, there is no defined standard methodology for rigging a face. Studios continue to redefine the techniques, processes, technologies and production pipelines to efficiently create films and videogames.

Today, facial animation is done manually by skilled artists, who carefully place and manipulate the animation controls to create the desired motion. As models become more and more complex, it is increasingly difficult to define a consistent rig that can work well for many different characters. So each facial rig has to be created individually by hand. This traditional method ensures high quality results, but it is slow and costly. Large film and videogame companies can afford hiring lots of artists, but this is not feasible for low budget production. It takes an experienced digital artist from one to four weeks to create a complete facial rig, depending on its complexity. But if any change must be applied to an already created rig , the rigging process has to restart. Facial rigging becomes a serious bottleneck in any CG production.

Finding the optimal solution to create a facial rig depends on several constraints: time to develop the rig, budget, artists experience, expected rig performance and actions, and others. The three most common approaches to create a rig are based on: blend shapes [17], bones [28] or a combination of both [16]. However, there are other existing facial animation methods, like motion capture, that can produce photorealistic results and speed up the animation process, but are unable to adapt the performance to dissimilar characters. The captured animation will look the same in all models, ignoring their different appearances. Motion capture focus on analyzing what data to transfer, while our approach focus on what data to transfer and how to represent it.

Thus, the uniqueness of faces makes facial synthesis so challenging. The smallest anomaly in the face shape, proportion, skin texture or movement is immediately detected and classified as incorrect. Most rigging challenges are:

- **no standard:** artists do not follow a formal criteria or methodology when creating a rig, making it difficult to create a solid platform to build upon;

- **changing the geometry or resolution:** it is very common to change the face model during production, to improve the deformation details or simply because it looks better. Any minor modification in the model surface (a bigger nose, more resolution around the lips) after the character is rigged, causes the rigging process to restart;

- **reusing weight maps:** the weight distribution defined for one character will not work on others.

- **number of shapes leads to complex user interface:** many productions use rigs based on hundreds of shapes. Usually, too many shapes make it hard to use the rig. Likewise, if a shape is added during production it can generate two problems: the shape conflicts with existing animations, making it necessary to rework some shots; or the new shape does not mix nicely with the the others;

- **preserving a consistent look:** placing by hand the animation controls leads to different artistic interpretations of where to position each element of the rig. This makes it difficult to easily reproduce the same facial pose between different characters. Consequently, it becomes hard to guarantee a consistent look throughout the production;

4. Our Facial Rigging System

Creating and placing by hand each component of the rig (bones, controls) quickly becomes impractical when complexity grows. The system we present can handle simple and complex rigs based on a new approach described in our previous work [19]. The system is:

- **generic:** the facial rig can have any type of configuration and does not force the use of a predefined rig;

- **flexible:** the rig has no initial constraints;

- **independent of the shape:** a facial rig can be transferred between models that have different geometry, look and appearance;

- **enhances artistic freedom:** artists can use any tool or deformation method to create the rig.

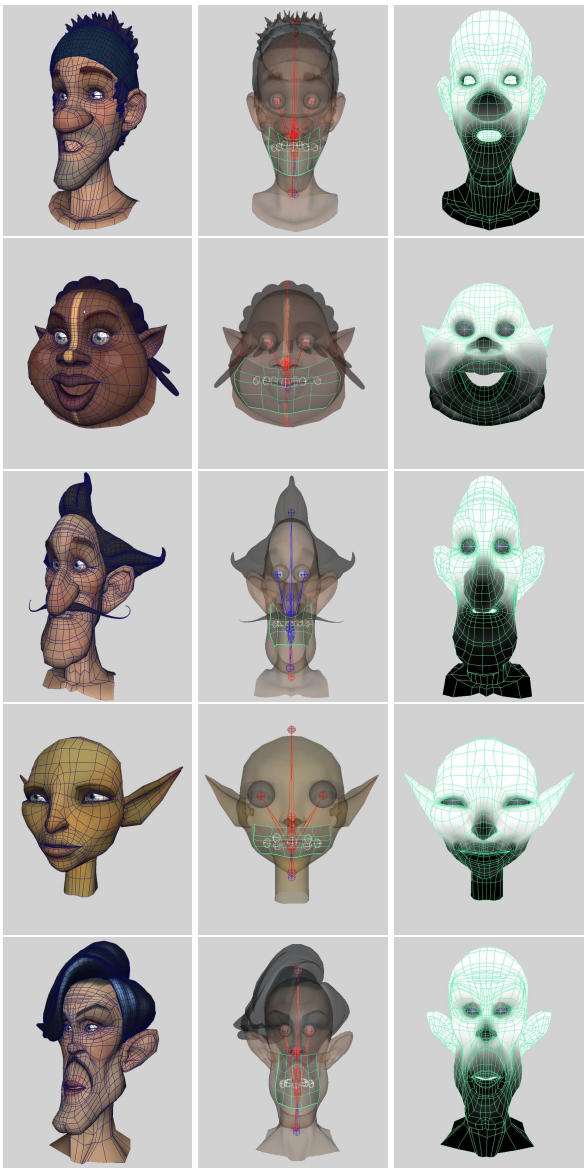


Figura 2. The first row shows the source model (Lisandro) and the rest show the target models (Mostaza, Teseo, Hada and Demetrio); first column shows the look and appearance of the models; second column details the facial rig that includes 21 joints and 1 NURBS surface; and third column shows the weight distribution. All models have different wireframe. (Copyright 2005 Dygrafilms)

4.1. System Description

The system deals with the setup of the character rig. It allows using the rig created for one character in others. To transfer the facial rig we begin by defining two 3D face models. The first one, we call source model, is rigged and includes a series of attributes: a control skeleton, a number of influence objects that represent the inner structure of the face and animation controls, facial expressions (shapes) and animation scripts. The rig doesn't have initial constraints

and can be created by an artist. The second model, we call target model, doesn't have a character rig associated to it. The source and target models can have different descriptors: one can be defined as a polygonal surface and the other as a NURBS surface. Also, the faces do not need to have the same number of vertices. Figure 1 shows an overview of the system pipeline and illustrates the rig transfer process with two dissimilar characters. The main steps within the system are: 1. surface deformation; 2. attribute transfer; 3. skinning.

1. Surface Deformation The source rig information is used as the direct input for transferring the setup to the target model. First, our deformation method deforms the source model surface to match the geometry of the target. We landmark the facial features to keep correspondence between source and target model, and then employ a computer vision interpolation technique named Thin Plate Splines (TPS) [3], as our deformation kernel function. After the TPS, the source surface only has exact deformation at the landmark positions of the target model, while the rest of the points lay outside the target surface. We solve this by applying a dense correspondence algorithm [19], which projects every point of the warped surface to the closest point of the target and determines the correspondence between every source and target vertex.

2. Attribute Transfer Using as reference the previously deformed source surface, we call guide model, the method accurately places the source rig attributes (section 4.2 describes the rig attributes) into the target model, even if they have different geometric proportions. We had to adapt the TPS to properly deal with each attribute specific characteristics. It is not the same transferring bones than transferring a NURBS curve. The dense correspondence avoids placing additional landmarks on the influence objects or on the skeleton structure. The deformation process achieves excellent results in positioning the source rig attributes in the correct regions of the target face. For example, joints and NURBS surfaces are relocated in the target model, based on the correspondent position they have in the source model. They are also transformed to fit the shape and size of the target (see figure 2).

3. Skinning After the deformation step comes the skinning, based on a smooth binding algorithm. It binds the transferred attributes to the target model using the adjusted weights of the source, avoiding the need for manual weighting. The weights at the target are calculated using the deformation method. Each vertex of the target model accurately adapts the blending weight of the joints and influence object, based on the source model weight distribution, to properly represent the target facial look and behavior (see figure 2). Last, as the target model is already rigged and weighted, transferring facial animations is a straightforward process. The method only needs to scale and adapt the animation curves to fit the proportions of the target. The end result are face models ready to be animated with production quality rigs.

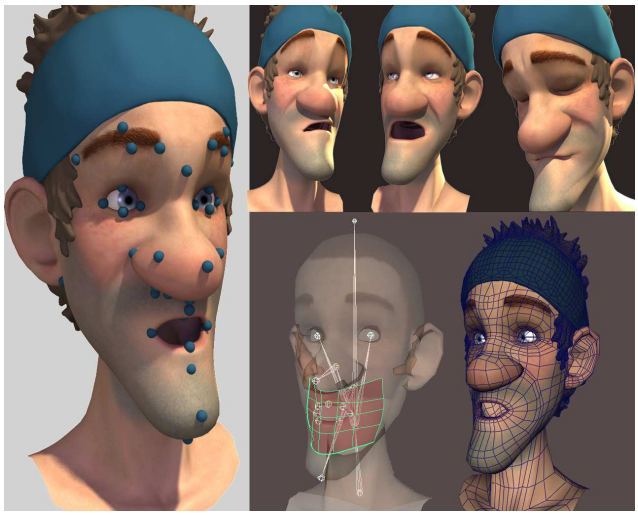


Figura 3. Source rig used in our examples; images show different rig attributes: landmarks, expressions (created using the rig and shapes), joints and NURBS surface, wireframe. (Copyright 2005 Dygrafilms)

4.2. Rig Definition

Central to our system is the notion of source rig $\mathcal{S}$, and we use the model in figure 3 to illustrate it. The rig is formed by different layers of abstraction that we refer to as attributes: skin surface $\mathcal{S}_S$, influence objects $\mathcal{S}_O$, skeleton bones $\mathcal{S}_B$, facial features landmarks λ, shapes $\mathcal{S}_H$, animation scripts $\mathcal{S}_A$ and other components for representing the eyes, teeth and tongue. We can assign additional attributes to each of these layers: weight, texture, etc. [10].

The source rig helps define the appearance of the characters. It establishes the character setup standard shared by all the models. Artists can create their own source rig, because they are free to use any type of controls and components to achieve the desired visual look.

The **source rig** $\mathcal{S}$ has been modeled manually and is a highly deformable structure of a face. During the modeling process, we used facial features and regions to guarantee realistic animation and reduce artifacts.

The **surface** $\mathcal{S}_S$ is the external geometry of the character that determines the skin of the face, using polygonal surfaces composed by a set of vertices $\mathbf{r}$ and a topology that connects them.

The source rig is tagged with **landmarks** λ, distributed as a set of sparse anthropometric points. We use the landmarks to define specific facial features to guarantee correspondence between models.

The **skeleton** $\mathcal{S}_B$ is a group of bones positioned under

the skin. It defines the pose of the head and controls lower level surface deformation. Each bone is defined by two joints, one at each end of the bone.

The **influence objects** $\mathcal{S}_O$ are objects that affect the shape of the skin and help artists control the 3D models. They include: NURBS surfaces, NURBS curves, lattice deformers, cluster deformers, polygon mesh, and others.

The **shapes** $\mathcal{S}_H$ are new 3D face models created by applying deformations over the geometry $\mathcal{S}_S$ of the character. A shape is a 3D facial pose of the source model, where $\mathcal{S}_H$ and $\mathcal{S}_S$ have the same geometry. Shapes are usually modeled manually by an artist. They represent facial expressions or partial deformation of a specific area of the face. They are used to create blend shapes, which let you change the shape of one object into the shapes of other objects. The interpolation between shapes results in facial animations.

The **animation scripts** $\mathcal{S}_A$ consist of a list of animation curves that determine motion. Each animation curve represents changes in the value of an attribute, like shapes or bones.

4.3. Application and Workflow

We implemented a set of plug-ins in C++ for Maya [1]. The plug-in includes a simple user interface to ease the landmarking and assist the transfer process (see figure 4). The modular design of the application makes it simple to integrate into existing animation pipelines.

The application enables artists to fit automatically the rig from the source to the target model; manipulate the target as if they were using a puppet; and adjust animation parameters in the target model or animate the target using predefined source animations.

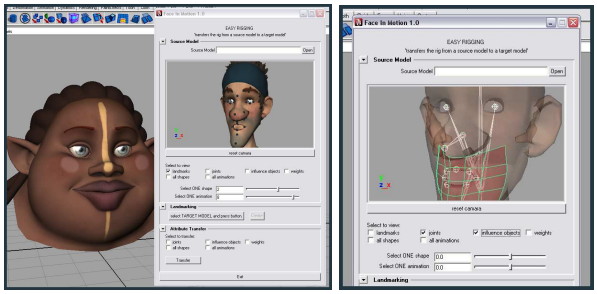


Figura 4. Application user interface running on Maya: assisting the landmarking process (left); close up of the source rig viewer with joints and influence object selected (right). (Models copyright 2005 Dygrafilms)

The *input* to the pipeline is the source model $\mathcal{S}$ information. The *output* is a fully rigged target model $\mathcal{F}$ ready to be animated. The workflow of the application is as follows:

1. **Landmarking**: Defines the source and target model landmarks that will keep correspondence between models.
2. **Surface Correspondence:** Ensures the exact point matching at the landmarks and smoothly interpolates the deformation of other points.
3. **Surface Dense Correspondence:** Ensuring exact deformation of every surface point, avoids placing additional landmarks.
4. **Attribute Transfer:** Uses the TPS deformation method to transfer each type of attribute.
5. **Skinning:** Binds the deformable objects, influence objects and surface to the skeleton of the target model.

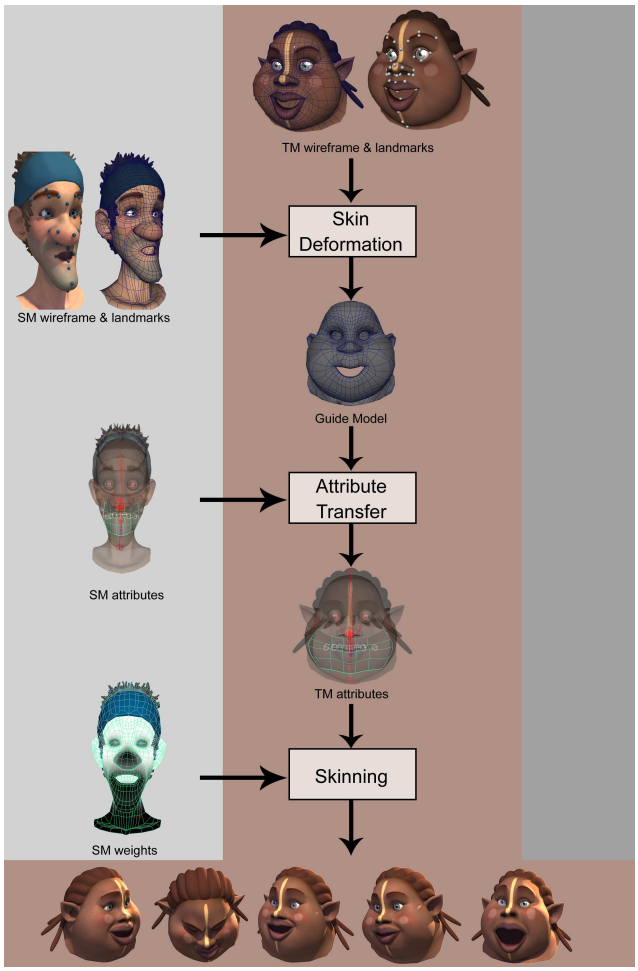


Figura 5. System pipeline. Shows the three main steps needed to transfer a facial rig: skin deformation, attribute transfer and skinning. The output of the skin deformation is the guide model, which serves as reference for the rest of the transfer process. The output of the attribute transfer is the target model with the rig components positioned in correspondence to the source. Last, after skinning the character using the source model weights, the target model is rigged and ready to be animated. (Copyright 2005 Dygrafilms)

5. Results and Discussion

Reproducing the subtleties of a face through animation requires developing a sophisticated character rig. But, creating by hand the inner structure and controls of each character is a very labor-intensive and time-consuming task. We presented a system that transfers the rig and animations between characters, at least an order of magnitude faster than traditional manual rigging. The system allows *creating* the rig, animation controls and scripts for *one model* (source), and *reuse* them in many different *target models*. It is independent of the appearance or shape of the model, so rig transfer between dissimilar characters is feasible. Artists can create their own rigs and are not forced to use predefined ones. In film and videogame productions, artists are often given one base model to make all new faces (shapes). Also, it is common that afterwards they are asked to use a different 3D face, because it has improved deformation details or simply looks better. Currently, all shapes need to be remade to reflect the topology of the new face. But our method makes sure that previous work can be transferred and artists time is not wasted.

Our facial animation system can be integrated into existing animation production pipelines, improving its work flow as it decouples the work of the animators and the modelers, they can be working in parallel on the same character. As a result, the companies will have fewer bottlenecks, which will increase productivity and reduce cost.

The system also provides a solid foundation for setting up consistent rigging strategies: at the beginning of a production, artists can define the required rig parameters and afterwards use them as a template for all models. This rig becomes the building block for all characters. Our approach helps film and videogame studios overcome the current lack of a standard rigging methodology. It guarantees that all rigs generated by the system produce homogeneous results, ensuring that the models share a common vision and consistent artistic style.

Testing and Validation We validated the system with a series of experiments. We used source models from several companies (Electronic Arts, Blur Studios, Dygrafilms, etc.) and for each, we transferred the rig and animations to different target models. We worked with a variety of styles: human, cartoon and fantastic creatures. Then, for the same models, we compared the output of our application with the results manually created by an artist. The results were supervised by Technical and Art Directors, who approved the quality of our rig and animations to be used in CG productions, replacing the artist generated ones (see figure 6 for a detail explanation). This is a crucial result: if the output still requires a lot of tuning, then the system is useless in a production. The examples of the paper are limited to synthetic characters to emphasize the versatility of the method.

Performance Our application allows creating the rig in one hour as we need to visually validate the results, (go through all the rig) which take some time if the rig is very complex. The attribute transfer process like changing the weights, modifying a control position or transferring animations, is nearly instantaneous. Figure 7 shows that our method convincingly captures the complex effect of simulating a talking head, to be used in a film. The big time savings achieved on the rigging process is usually an order of magnitude or more, and still meet the high quality animation needs of the entertainment industry. The tests were made on a AMD Athlon 64 3500+ CPU with 2 GB of RAM.

Extreme Test To test the method to the extreme of its

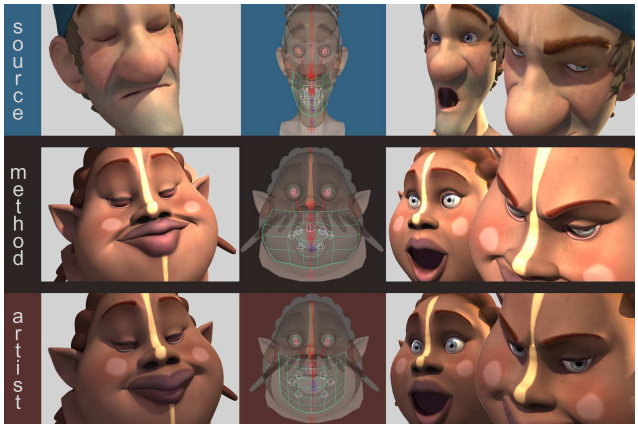


Figura 6. Comparison of the target models automatically generated by the method and manually created by an artist. The images on the middle show a NURBS surface that simulates the orbicularis mouth muscle: we can see that the method is more precise than the artist when adapting it to the target geometry. This is a particulary difficult task for the artist; he has to make sure that the NURBS parameters are homogeneous to avoid strange deformations or artifacts in the mouth (like pinching in the lips). This is a also an example of transferring animations between dissimilar geometries: the method preserves the intensity of the facial expressions and animations of the source. (Copyright 2005 Dygrafilms)

Figura 7. Source (1st row) and target models (2nd and 3rd row) animation sequence: frames from a video that integrates facial animation with lyp-sync. (Copyright 2005 Dygrafilms)

possibilities, we ordered three very different 3D models: a photorealistic human (source model), a cartoon and a fantastic creature (target models). The models differ enormously in artistic style, deformation behavior, shape and proportions (see figure 8). The source model rig includes: 2 NURBS curves around the mouth, 1 lattice for the jaw, 6 joints for the head, 5 joints for the tongue, 3 joints for the teeth, 47 shapes and 2 animations clips (one with extreme facial poses and the other with lip-sync). The method successfully transferred the shapes in the mouth region, which is very complex due to the variety of poses it can perform. But the drawback of transferring shapes between characters with different styles is that the target models inherit the movements of the source. We obtained a cartoon model simulating a human character. This faithfulness it is not always what the artist wants, so we needed to keep in mind this behavior.

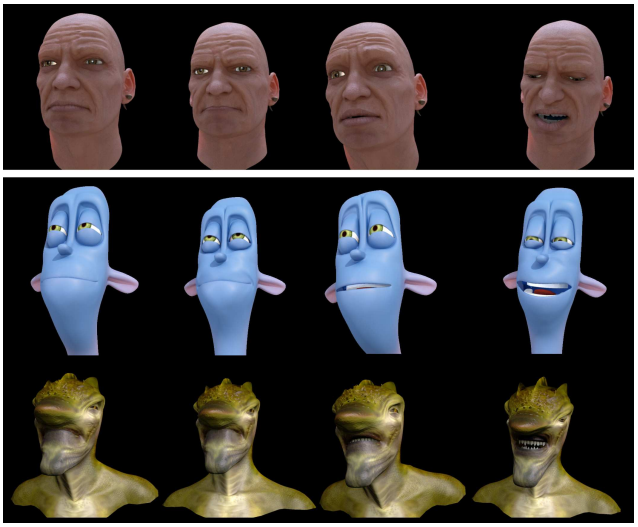


Figura 8. Source (top) and target models (bottom). Keyframes extracted from a video sequence to show different poses transferred from the source to the target models. The poses were created by manipulating the source rig. (Copyright 2007 Face In Motion)

Limitations An important issue to mention is that if the source rig quality is low, the transference is still successful but the results on the target will be of comparable quality. The technology is indeed independent of the quality and the shape of the rig. During our tests, we realized that when the source model has the eyes and the mouth completely closed, the attribute transfer results show some artifacts. This a current limitation of our solution. To obtain artifact free transfers, it is recommended to have the eyes and mouth of the source and target models slightly opened.

Future Work We performed some tests on mapping motion capture data into the source rig, and later transfer it to the target model. This is an interesting direction for future research and to extend our application.

6. Acknowledgement

Many thanks to Juan Nouche (Ottiplanet) and Xenxo Alvarez (Enne Studios) for their feedback and for providing the 3D models while working at Dygrafilms. Also to Fred Fowels (Rainmaker), Jean Luc-Duprat (Intel) and Crystal Wang (Electronic Arts) for testing the system and providing 3D models and motion capture data. Special thanks to Toni Susin for supervising the research project. Last, we would like to thank Blur Studios for using our system into their CG pipeline to create the facial rig of the Simpsons and Fable productions. This project is partially funded by FCT, Portugal (http://www.it.pt).

Referências

[1] I. AUTODESK. Autodesk maya, 2008. http://www.autodesk.com/maya. 2, 5

[2] I. Baran and J. Popović. Automatic rigging and animation of 3d characters. In *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, page 72, New York, NY, USA, 2007. ACM. 2

[3] F. L. Bookstain. Principal warps: Thin-plate splines and the decomposition of deformations, 1989. IEEE Trans. Pattern Anal. Mach. Intell. 11, 6, 567585. 4

[4] G. Borshukov, J. Montgomery, and W. Werner. Playable universal capture: compression and real-time sequencing of image-based facial animation, 2006. In SIGGRAPH'06. 1, 2

[5] J. E. Chadwick, D. R. Haumann, and R. E. Parent. Layered construction for deformable animated characters. *SIGGRAPH Comput. Graph.*, 23(3):243–252, 1989. 2

[6] J. X. Chai, J. Xiao, and J. Hodgins. Visionbased control of 3d facial animation, 2003. In SCA'03. 2

[7] Z. Deng, P. Chiang, P. Fox, and U. Neumann. Animating blendshape faces by cross-mapping motion capture data, 2006. ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3DG06). 1, 2

[8] R. Falk, D. Minter, C. Vernon, G. Aretos, L. Modesto, A. Lamorlette, N. Walker, T. Cheung, J. Rentel-Lavin, and H. Max. Art-directed technology: Anatomy of a shrek 2 sequence, 2004. ACM SIGGRAPH'04 Course Notes, ACM Press, NY, USA. 2

[9] D. Fidaleo, J. Y. Noh, T. Kim, R. Enciso, and U. Newumann. Classification and volume morphing for performance-driven facial animation, 2000. In Int. Workshop on Digital and Computational Video. 2

[10] J. Haber and D. Terzopoulos. Facial modeling and animation. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Course Notes*, page 6, New York, NY, USA, 2004. ACM. 4

[11] C. Hecker, B. Raabe, J. Maynard, and K. V. Prooijen. Real-time motion retargeting to highly varied user created morphologies, 2008. In SIGGRAPH'08. 2

[12] W. M. Hsu, J. F. Hughes, and H. Kaufman. Direct manipulation of free-form deformations. *SIGGRAPH Comput. Graph.*, 26(2):177–184, 1992. 2

[13] P. Joshi, W. Tien, M. Desbrun, and F. Pighin. Learning controls for blend shape based realistic facial animation, 2003. In SCA'03. 2

[14] K. Kahler, J. Haber, H. Yamauchi, and H. P. SEIDEL. Head shop: generating animated head models with anatomical structure, 2002. In SCA'02. 2

[15] Y. Lee, D. Terzopoulos, and K. Walters. Realistic modeling for facial animation. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62, New York, NY, USA, 1995. ACM. 2

[16] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2

[17] C. Maraffi. Maya character cretion: Modeling and animation controls, 2003. New Riders Publishing. 2

[18] J. Noh. *Facial animation by expression cloning*. PhD thesis, Los Angeles, CA, USA, 2002. Adviser-Ulrich Neumann. 1

[19] V. C. Orvalho. *Reusable Facial Rigging and Animation: Create Once, Use Many*. PhD thesis, Barcelona, Spain, 2007. Adviser-Antonio Susin. 2, 3, 4

[20] F. Pighin and J. P. Lewis. Facial motion retargeting, 2006. In SIGGRAPH06: ACM SIGGRAPH Courses, ACM Press, New York, NY,USA. 1

[21] K. Richie, O. Alexander, and K. Biri. The art of rigging vol. 2, 2005. CG Toolkit. 1

[22] J. Schleifer. Character setup from rig mechanics to skin deformations: A practical approach, 2004. ACM SIGGRAPH'02 Course Notes, ACM Press, NY, USA. 2

[23] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. In *SIGGRAPH '86: Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 151–160, New York, NY, USA, 1986. ACM. 2

[24] E. Sifakis, I. Neverov, and R. Fedkiw. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Trans. Graph.*, 24(3):417–425, 2005. 2

[25] K. Singh and E. L. Fiume. Wires: a geometric deformation technique, 1998. In SIGGRAPH'98. 2

[26] R. Turner and D. Thalmann. The elastic surface layer model for animated character construction, 1993. In Computer Graphics International'93. 2

[27] K. Wampler, D. Sasaki, L. Zhang, and Z. Popović. Dynamic, expressive speech animation from a single mesh. In *SCA '07: Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 53–62, Aire-la-Ville, Switzerland, Switzerland, 2007. Eurographics Association. 1

[28] A. Ward. Game character development with maya, 2004. New Riders Publishing. 2

Motivations, Strategies, and Movement Patterns of Video Gamers Playing Nintendo Wii Boxing

Marco Pasch[1], Nadia Berthouze[2], Betsy van Dijk[1], Anton Nijholt[1]

[1]Human Media Interaction,
University of Twente, The Netherlands
{m.pasch, bvdijk, anijholt}@ewi.utwente.nl

[2]UCL Interaction Centre,
University College London, UK
n.berthouze@ucl.ac.uk

Abstract

Video game consoles that employ physical activity as an interaction mode can benefit from using the gamer's movement as feedback and adapt to it. But to be able to design such systems we need to know how gamers actually move and what we can infer from this. This paper reports preliminary, qualitative results of a study that aims at identifying playing styles and related movement patterns of gamers that play the Nintendo Wii Boxing game. Interviews of video gamers revealed that they approach the game with two different motivations (to achieve and to relax) that lead to different strategies (game and simulation). A movement analysis study using motion capture data, video recordings, and observer ratings identified three different movement patterns that relate to these strategies.

1. Introduction

A new generation of video game consoles enables video gamers to employ active body movements as interaction mode. Initial studies show that usage of such movement-based consoles reaches exertion levels that increase physical health and appear promising for reducing obesity [10], [11], which is at least partly the result of a sedentary lifestyle [12].

Presently, there is a fair amount of research done to develop intelligent interfaces [16], i.e. interfaces that employ the user's behaviour as input and adapt to it. Such interfaces usually act on affective cues that are inferred from the users. But with the advent of movement-based interaction, intelligent interfaces also have to analyze the movement patterns of their users. For instance, what does the force when swinging a baseball bat or the number of punches when boxing say about the gamer's experience? How can we detect a gamer over pacing himself or moving in an unhealthy way?

Apart from the challenge of developing technology that can detect movement patterns of the user, another important task - and one that should precede the development of technology - is to investigate and identify the movement patterns that users display. If we want to build systems that react on the user's behaviour we must know what we can observe and what it means in the first place. This is not a trivial task as the human body is complex with a large number of degrees of freedom,. These are usually represented by joint rotations (e.g. around the shoulder joint or elbow joint) to describe the movement that one carries out.

Some approaches to intelligent, movement-based interfaces circumvent joint rotations and just consider the amount of movement in an image or track the head or the hands. But also here the task of identifying patterns in the movement and what they mean remains.

It is the goal of this paper to contribute to the search for behavioural patterns and their meanings. In the study described here we attempt to identify movement patterns of video gamers that are playing on the Nintendo Wii. The Wii is a popular video game console that gamers steer by using one or two handheld controllers that are fitted with accelerometers and that allow the console to detect the location of the controllers in 3D space.

In a first step we conducted interviews with video gamers to investigate how they conceptualize and interpret their movements when playing movement-based games. This revealed two different motivations and corresponding strategies for playing. In a second step we investigated if the two strategies can be found back in movement patterns that gamers exhibit while playing. For this, we fit gamers with an inertial gyroscopic motion capture suit and observed their movements while they played the Wii Sports Boxing game. The Wii Boxing game was chosen because the Wii is presumably the most widely distributed movement-based video game console at present and Wii Boxing reaches the highest activity levels of the Wii Sports games [10].

The paper is organized as follows: We first discuss potential benefits of movement-based games as opposed to sedentary games and outline areas where adaptive interfaces could improve the interaction. We then present the outcome of the interviews with video gamers. The movement analysis study is described next. Finally, we discuss the results of both studies and give pointers to future research.

2. Potential Benefits of Intelligent, Movement-based Video Games

A video game that employs both active body movements as interaction mode and can adapt to the gamer's behaviour, offers several potential benefits. Before moving on to our findings, we want to discuss two domains of potential benefits, arising from such games: (a.) a healthier interaction, and (b.) a richer, more enjoyable interaction.

2.1. Healthier Interaction

Video games are usually seen as contributors to the growing obesity epidemic [5]. Hillier [12] notes that "children today are engaging much less with the world outside their homes in terms of physical activity ... Technological innovations in media have contributed to these changes, keeping children inside and sedentary in their playtime..." (p. 56). Yet, instead of simply blaming technology, she advocates making technology part of the solution. Physical activity promoting video games can be seen as an example of such a technology driven solution.

Initial studies show that physical activity during gameplay increases energy expenditure significantly compared to sedentary games. Lanningham-Foster and colleagues [14] measured the energy expenditure of children playing sedentary video games and playing active video games like Sony's EyeToy and Konami's Dance Dance Revolution. The energy expenditure more than doubles for Dance Dance Revolution and the authors conclude that such games could be useful for obesity prevention and treatment.

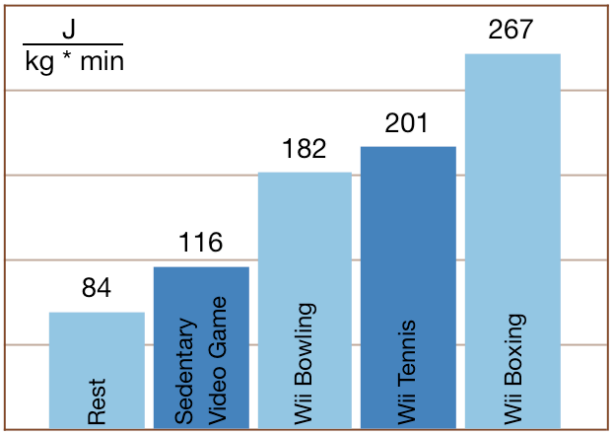


Fig. 1: Energy Expenditure of Adolescents whilst Playing Video Games, from Graves and colleagues [10]

Graves and colleagues [11] also measured the energy expenditure of children playing active video games as compared to sedentary games, but observed an older group of children. Interestingly, they compare the expenditure values of the Wii Sports Bowling and Tennis with values for real Bowling and Tennis. The Wii games require significantly more energy than sedentary activities, but less than the real sports that they simulate. Yet, the measure they employ does not take upper limb movements into account. In a more recent study Graves and colleagues [10] use a measure for estimating the energy that includes upper limb movements, which form a crucial part of interacting with the handheld controllers of the Wii and that were neglected in [11]. Including upper limb movements promises more accurate results and indeed, they report higher activity levels than in the previous study. Figure 1 shows the energy expenditure levels that Graves and colleagues [10] found for the Wii Sports games, in comparison to a rest value and a sedentary video game on the Microsoft XBOX 360 video game console.

Of all the Wii Sports games (i.e. baseball, bowling, boxing, golf, and tennis) boxing reaches the highest activity levels. The authors conclude that while the intensity of real boxing is much higher, the intensity of the Wii Boxing game exceeds the cut-off for moderate intensity physical activity. It is thus high enough to contribute to recommended amounts of exercise.

Besides in obesity prevention, active video games have also been envisioned for use in rehabilitation. The field of Virtual Rehabilitation has used virtual reality technology for some time now for the rehabilitation of patients [3][13], [17]. Recently, also physical activity promoting video game consoles have been deployed in rehabilitation measures. Morrow and colleagues [15] present a rehabilitation system, which is based on Microsoft's XBOX. They advocate the use of entertainment technology for physical rehabilitation, mainly to reduce system costs. Galego and Simone [7] combined a Wii remote control and Second Life into a Virtual Rehabilitation system. They also point out the potential of such low cost rehabilitation approaches. Though there are no results of scientific evaluations available yet, therapists have already coined the term "Wiihabilitation" and report of increased motivation of their patients, who are often unmotivated to carry out the very repetitive limb movements common in rehabilitation [19].

All this gives evidence for the benefits of video games that require the gamer to be physically active as compared to sedentary games. Still, they also expose gamers to new threats: Injuries from playing the Nintendo Wii have been reported in popular media and physicians have already introduced the diagnosis "Wiitis" [2] or "Wii shoulder" [4]. Bonis [2] describes the condition as follows: "If a player gets too engrossed, he may 'play tennis' on the video screen for many hours. Unlike in the real sport, physical strength and endurance are not limiting factors" (p. 2431). It is also an example for how much the interaction with the Wii is dependent on arm movement. One could also speculate that a further reason for such

injuries is that gamers do not perceive their video game consoles as sport devices and consequently do not care about warming up before playing [17]. This is certainly an issue that should be addressed in future research as well as in future game design. Otherwise the health improving effect of the physical activity can degrade.

By enabling game technology to monitor body movement and movement patterns, the game could be adapted at run time in order to foster a more positive and personal experience by encouraging healthier body movement. Once a threat is identified the game can then steer the gamer towards a healthier behaviour. Also, an adaptive game could monitor the exertion level of a gamer and steer the gamer towards recommended exertion levels.

2.2. Richer Interaction

The second domain we discuss is the promotion of a richer and more enjoyable interaction. Riskind and Gotay [18] found that the sheer posture of persons has influence on their mental state. Subjects that were put in a hunched, threatened posture reported greater stress than subjects that were put in a relaxed posture. Fox [6] reviewed studies that investigate the influence of physical activity on mental well-being. He concludes that there is growing evidence that exercise increases mental well-being, largely through improved mood and self-perception. Returning to a video game context, Bianchi-Berthouze and colleagues [1] found evidence that body movements not only increase the gamers' level of engagement, but also have an influence on the way a gamer becomes engaged. Their results demonstrate that the controller itself plays a critical role in creating a more complete experience for the gamer.

Whether the increase in engagement in physically active environments is due to the actual physical activity or to a higher perceived level of control remains open for research. Yet, more knowledge is needed about the link between physical activity and engagement in order to develop adaptive games that steer the gamer's movements towards a more enjoyable interaction.

Movement patterns can for examples shed light on the affective states and motivations of the gamer. A game technology able to capture such information and exploit it to adapt the game would provide a more natural and richer experience that could facilitate a sense of presence. If a gamer becomes aware that the game is reacting to his or her body movement, this may motivate the gamer to further exploit this channel of communication. This would offer a much richer set of strategies for challenging the opponent or communicate with possible teammates.

3. Interviewing Video Gamers

Interviews with video gamers were held to investigate how they experience, conceptualize, and interpret their movements when playing movement-based games.

3.1. Setup

Four experienced video gamers were recruited for this study. It did not appear useful to recruit novices, as some level of exposure is required for interviewees to reflect on their experiences with movement-based games. Interview sessions were held in a semi-structured style and initial outcomes were used to update the interview guide for the following interviews. Before the interview, subjects were primed by a 20 minutes session of playing the Nintendo Wii Sports games, during which they were videotaped. Subjects were instructed to play a game with a slow pace (i.e. bowling, golf, baseball) and a game with a fast pace (i.e. boxing, tennis) with the idea of asking about differences between the games, i.e., how the amount of physical activity and the type of movement may affect their gaming experience.

The interviews were transcribed and analyzed using a Grounded Theory approach, a qualitative methodology developed by Glaser and Strauss [8]. Aside from the statements of the interviewees, also observational data in form of memos was used in the analysis, as recommended by Goulding [9].

Open coding was applied to the data, i.e. labels were assigned to the statements of the interviewees and the observations. Then, relations between the labels were identified and finally put into concepts.

3.2. Results

A concept that emerged early in the data was that gamers have several distinct motivations to engage with movement-based games. In fact, some experienced gamers seem to be aware of their changing motivation and adapt their gaming strategy accordingly:

"As you play and play you start to realize that you don't really need to swing and it's just a small movement that you need to make - so I tend to play more technically rather than emotionally. […] When I am playing to relax and I play baseball, I swing like I would with a real baseball bat. But if I am playing to beat somebody else then I do what I need to do to do the movements." (i3)

The statement of interviewee 3 shows he has realized that he does not need to swing his arm with force. For the Nintendo Wii it is sufficient to make a small movement from the wrist. The challenge is thus the timing of the movement. In fact, to achieve a higher score it is beneficial to only make small movements from the wrist, as this allows more precise control. Nevertheless, the interviewee states that sometimes he deliberately makes big, forceful movements, when his motivation is not to achieve a high score, but just to relax and immerse into the virtual environment.

Gamers seem to appreciate the reduced complexity of the Wii compared to a real sport: "*Playing tennis in real life is harder"* (i4). Yet, there were also statements that

Table 1: Results from Video Annotation, Motion Capture, and Observer Ratings

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10
Video Annotation Data										
Number of Punches*	74	39	93	31	53	19	18	20	129	46
Ratio Hit to Total Punches*	0.28	0.41	0.22	0.71	0.34	0.68	0.33	0.4	0.11	0.41
Ratio Executed to Total Punches*	0.47	0.74	0.35	0.84	0.57	0.84	0.72	0.6	0.27	0.54
Motion Capture Data										
Displacement Body Root (Hip)*	357	22	76	232	316	261	572	301	139	386
Angular Displacement of Elbows*	55637	6539	47717	37274	35786	9264	14516	8190	13296	15699
Punch Amplitude (X-Rot., in Deg.)	95	8	118	160	90	63	90	105	18	75
Observer Ratings										
Boxing Realism**	2.3	1.2	2.2	2.4	2.5	3.6	3.1	3.8	1.7	3.7

* accumulated over 20 seconds, sample taken from middle of gaming session
** scale: 1 (low) – 5 (high)

gamers felt exhausted after playing on the Wii. Further, physical fitness was hardly mentioned by the interviewees and only as nice byproduct, but not as a motivation to engage with the game.

3.3. Discussion

We can conclude that there are two different strategies that gamers employ when playing a movement-based game and that they derive from different motivations to play in the first place. In the first case, the gamer is playing a game with the motivation to challenge his/her ability to find the best way to make points and have fun. The aim is to win and to achieve something. The related strategy is thus to maximize all efforts towards achieving a high score.

In the second case, the motivation for playing is to relax by experiencing and/or challenging their movement skills like they would do in a sport situation. Relaxation here does not refer to physical relaxation, but rather a mental relaxation that derives from immersing into the game and imagining oneself as playing the actual sport, not just a video game. Gamers that want to relax in such a game employ a different strategy. Instead of optimizing their gameplay towards achieving a high score they rather simulate the actual sport, i.e. they do the same movements as they would in the actual sport or how they think a good player would execute the movement in the real sport.

4. Movement Analysis Study

We conducted a motion capture study to investigate whether different motivations for playing and therefore deviating strategies identified from the interviews can be found back in movement patterns.

4.1. Setup

10 participants (thereof 7 males; mean age: 26 yrs, SD: 2.6) were fitted with an inertial gyroscopic motion capture suit (Gypsy 6, Animazoo, Brighton, UK) and their movements were recorded while they played the Wii Sports Boxing game for 15 minutes. To avoid biasing the participants, the experimenter left the room during this period.

In addition to the motion capture data, video recordings were made from a frontal-lateral angle and from over the shoulder of the gamer, to be able to correlate movements to game events.

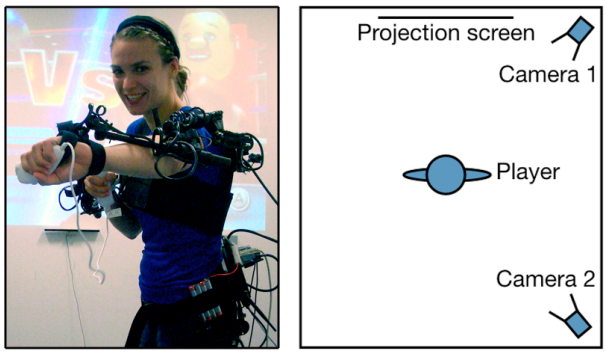


Fig. 2: Setup of the Motion Capture Study; Left: Participant in Gypsy 6 Motion Capture Suit; Right: Camera Positioning

Figure 2 shows the setup of the experiment. A third measure comes from five observers that rated video clips of the participants on boxing realism. Three 10-second video clips of each participant were shown in a random order to the observers. The results from all three measures are given in Table 1.

4.2. Results

A first analysis step was a visual inspection of the video footage. This revealed great differences in playing styles, i.e. differences in punch frequency, punch amplitude, and overall body movement. Roughly, three types of playing style were observed: One group of gamers only made very little extensions of the arms, while punching at a high frequency. Another group of gamers showed big

extensions of the arms and also punched at a high frequency, to the extent that it appeared they were over pacing themselves. In both groups it appeared that the gamers' behaviour was almost independent of game events, i.e. they showed only little defensive behaviour, even when their avatar was hit repeatedly.

The third group appeared to box realistically, i.e. with big arm extensions, a low to medium frequency of punches and reacting to game events.

In a next step we quantified the features that were deemed important during the visual inspection. The punch frequency was measured for each participant by annotating a short segment of the video recordings.

Figure 3 shows a representation of the punch frequency distinguishing between when a gamer punches but the Wii does not execute the punch in the game, when a gamer punches but misses the opponent ("executed + missed"), and when a gamer punches and hits the opponent ("executed + hit"). The Wii does not execute a punch e.g. when a punch is too soft or when a gamer punches while the avatar is still in the process of executing a previous punch or is recovering from being hit.

Aside from total punches, Table 1 also shows the ratio of punches that actually hit the opponent and the ratio of punches that are executed.

From the numerical data of the motion capture suit we obtained the total movement of the gamers, as the displacement of the body core over a period of 20 seconds. Another measure is the angular displacement of the elbows. This is an accumulation of the angular displacement (in arc degrees) of both elbows combined over a period of 20 seconds. Also, we obtained the average punch amplitude for each participant. Table 1 shows the punch amplitude for the X-Rotation, i.e. the extension of the arm in a forward direction. Figure 4 gives an example for a rotation around the X-Axis.

When plotting the observers' ratings of how much they thought the gamers are really boxing against the angular displacement of the elbows, we can easily identify three clusters that correspond to the playing styles that were mentioned above. Figure 5 shows that plot and in addition the average punch amplitude as bubble size.

The first cluster (P02, P09) only gets a low realism rating and is further characterized by low amounts of angular displacements and punch amplitudes. The second cluster (P01, P03, P04, P05) receives medium realism ratings, high angular displacements levels and big punch amplitudes. The group with the highest realism ratings (P06, P07, P08, P10) only shows low amounts of angular displacements. Yet, if we look at the size of the bubbles, we see that they show medium to large punch amplitudes. When looking at the video footage one can indeed observe that these gamers react to events that happen in the game, i.e. they wait for a good moment to punch the opponent and they also take a defensive stance while waiting.

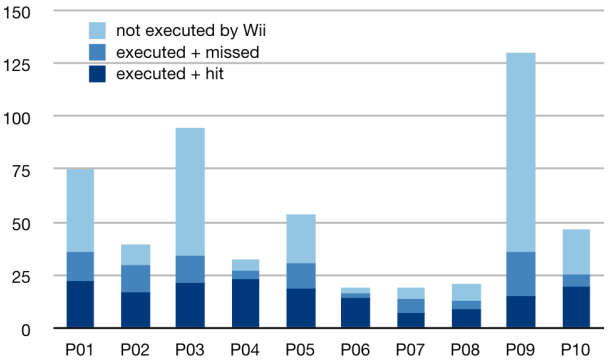


Fig. 3: Number of Punches over a period of 20 seconds.

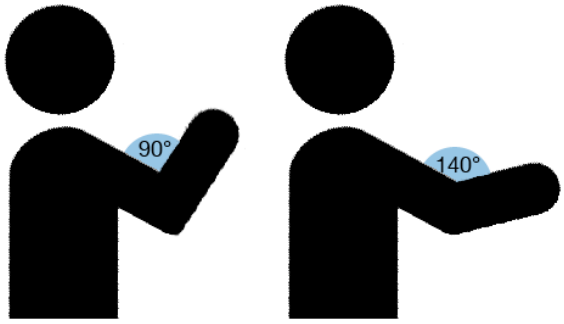


Fig. 4: Example of the punch amplitude for X-Rotation of the arms: A resting angle of 90° (left) and an extension angle of 140° (right) result in a punch amplitude of 50°

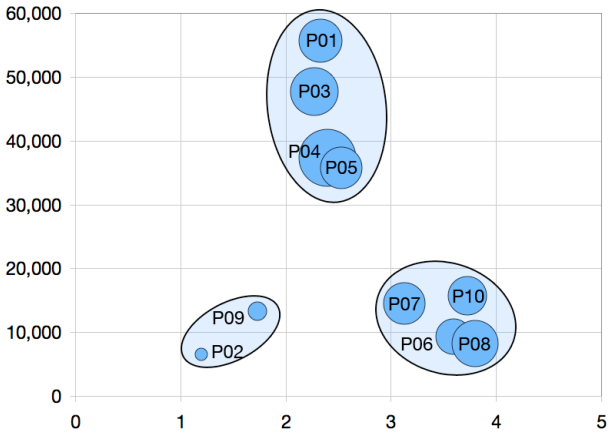


Fig. 5: Observer Ratings of Boxing Realism (X-Axis; Scale 1-5) vs. Angular Displacement of the Elbows (Y-Axis; accumulated over 20 seconds) vs. Average Punch Amplitude (Bubble Size; Size of X-Rotation, i.e. rotation in forward direction)

4.3. Discussion

In the interview study, we identified two motivations with which gamers approach the Wii Sports games: "Achieving" and "Relaxing". Also, corresponding

strategies were identified: “Game” and “Simulation”. In the movement analysis study we found three patterns, i.e. the clusters shown in Figure 5: One pattern corresponds to “Simulation”, while we have to differentiate for “Game”. Here, there appear to be two different patterns.

The first pattern can be described as “game with a low intensity”, i.e. gamers show only little physical engagement. The body core remains stationary and there are only very small arm extensions. Yet, they show a high punch frequency. Apparently these gamers have learned that for the Nintendo Wii the punch amplitude is irrelevant and that a short impulse is enough to perform a punch. The high punch frequency leads to a good performance in terms of total hits, even if many punches do not hit or are not executed at all. Still, the level of physical activity remains low and on the video recordings they do not appear to be emotionally engaged and almost look bored.

The second pattern can be described as “game with high intensity”. These gamers are quite active, i.e. they move around and show high arm extensions and punch amplitudes. The punch frequency varies from medium to high.

The pattern “simulation” is characterized by gamers that observe the action on the screen and react to it. They have a lower number of punches as they wait their turn and do not punch blindly. On the other hand they show big arm extensions, as is done in real boxing, which they simulate.

Figure 6 gives an overview of the motivations, strategies, and movement patterns that were identified for video gamers playing Wii Boxing.

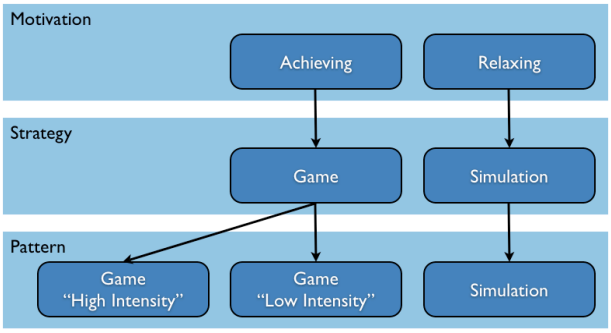


Fig. 6: Motivations, Strategies, and Movement Patterns of Video Gamers Playing Nintendo Wii Boxing

The following features appear thus important for identifying the gamer’s movement behaviour. The average punch amplitude gives a good indicator to distinguish between the low intensity pattern of “Game” and the pattern “Simulation”. Yet, it seems not sufficient to distinguish the high intensity pattern of “Game” from “Simulation”. For this, the angular displacement of the elbows appears suited (see Figure 5).

5. Conclusions

The aim of this paper was to identify playing styles and corresponding movement patterns for gamers of physically active games, in this case the boxing game of the Nintendo Wii Sports games.

We identified two motivations with which gamers approach the Wii Boxing game (“Achieving” and “Relaxing”) and two related strategies they employ (“Game” and “Simulation”). In the first one (“Game”) gamers are aiming for a high score and to achieve this reduce their movements to what is necessary. This can result in two different movement patterns, one with low punch amplitude and corresponding low physical intensity and one with a high punch amplitude and high physical intensity. In common for both patterns is that gamers punch at a high frequency and neglect events in the gameplay like their avatar being hit.

Gamers that want to relax and to immerse into the game use a different strategy, which we call “Simulation”. In the corresponding movement pattern they appear to imitate real-life boxing, i.e. they observe the opponent, try to block its punches and wait for good opportunities to attack.

The significance of the findings reported in this paper must be qualified by the rather small size of the pool of subjects. Still, our results identify trends and can help reduce the complexity of information that we obtain from movement data.

In a next step, the features that we identified here should be validated in a quantitative study. Also, other game scenarios should be investigated.

Reports from the interview study lead us to speculate of changes of the gamers’ behaviour as they gain expertise in a game. We also found that a gamer can approach a game with changing moods and motivations. A longitudinal study could investigate how motivations and movement patterns change over time and exposure.

Another aspect of this new type of physically active games is a social one. From our interviews we learned that for this type of game, gamers appear to meet with friends to play, as a sort of social event. Interviewees reported of a preference to play in a social setting. We did not consider social aspects in this study, but they should be addressed in future research.

As in all game research, a further critical issue is the artificial setting of laboratory studies. This makes it hard to get a good and reliable measure for the gamers’ experience. The use of a motion capture suit in this study was also quite intrusive and potentially influences the gamers’ experience. The identification of relevant features should limit the amount of necessary technology to record movement of the gamer and help towards designing future studies into a more natural setting.

All this should enable us to inform the design of user-

adaptive active games that steer the gamer towards a healthier and richer interaction.

Acknowledgements

The research of the authors of the University of Twente 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). The research of Nadia Berthouze has been supported by the Marie Curie International Re-Integration Grant "AffectME" (MIRG-CT-2006-046434).

References

[1] Bianchi-Berthouze, N., Kim, W.W., Darshak, P. (2007). Does Body Movement Engage You More in Digital Game Play? and Why? *Affective Computing and Intelligent Interaction*, Springer, LNCS 4738, 102-113

[2] Bonis, J. (2007). Acute wiiitis, *N Engl J Med*, 356(23), 2431-2432.

[3] Burdea, G. C. (2003). Virtual rehabilitation - benefits and challenges, *Methods Inf Med*, 42(5), 519-523.

[4] Cowley, A. D., & Minnaar, G. (2008). Watch out for wii shoulder, *BMJ*, 336, 110.

[5] Epstein, L. H., Roemmich, J. N., Robinson, J. L., Paluch, R. A., Winiewicz, D. D., Fuerch, J. H., et al. (2008). A randomized trial of the effects of reducing television viewing and computer use on body mass index in young children, *Archives of Pediatrics and Adolescent Medicine*, 162(3), 239-245.

[6] Fox, K. R. (1999). The influence of physical activity on mental well-being, *Public Health Nutrition*, 2(3a), 411-418.

[7] Galego, B., & Simone, L. (2007). Leveraging online virtual worlds for upper extremity rehabilitation. *NEBC '07*. IEEE 33rd Annual Northeast Bioengineering Conference (pp. 267-268).

[8] Glaser, B., Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York, Aldine

[9] Goulding, C. (2002). *Grounded Theory: A Practical Guide for Management, Business and Market Researchers*, SAGE, London

[10] Graves, L.E., Ridgers, N.D., Stratton, G. (2008). The contribution of upper limb and total body movement to adolescents' energy expenditure whilst playing Nintendo Wii. *Eur. J. Appl. Physiol*.

[11] Graves, L., Stratton, G., Ridgers, N. D., & Cable, N. T. (2007). Comparison of energy expenditure in adolescents when playing new generation and sedentary computer games: cross sectional study, *BMJ*, 335, 1282-1284.

[12] Hillier, A. (2008). Childhood overweight and the built environment: making technology part of the solution rather than part of the problem, *The ANNALS of the American Academy of Political and Social Science*, 615(1), 56-82.

[13] Holden, M. K. (2005). Virtual environments for motor rehabilitation: review, *Cyberpsychol Behav*, 8(3), 187-211.

[14] Lanningham-Foster, L., Jensen, T. B., Foster, R. C., Redmond, A. B., Walker, B. A., Heinz, D., et al. (2006). Energy expenditure of sedentary screen time compared with active screen time for children, *Pediatrics*, 118(6), 1831-1835.

[15] Morrow, K., Docan, C., Burdea, G., & Merians, A. (2006). Low-cost virtual rehabilitation of the hand for patients post-stroke. *Int. Workshop Virtual Rehabilitation,* 6-10

[16] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2007). Human Computing and Machine Understanding of Human Behavior: A Survey. *Artificial Intelligence for Human Computing, LNCS 4451, 47-71*

[17] Powell, V. (2008). CHI 2008 Exertion Interfaces: A flexible approach.. Paper presented at the *ACM CHI 2008 Exertion Interfaces Workshop*, May 2008.

[18] Riskind, J. H., & Gotay, C. C. (1982). Physical posture: could it have regulatory or feedback effects on motivation and emotion?, *Motivation and Emotion*, 6(3), 273-298.

[19] Tanner, L. (2008). Break a leg? Try 'Wiihabilitation'. *msnbc*, Retrieved on 18.06.08 from http://www.msnbc.msn.com/id/23070190/

Virtual Mirror Gaming in Libraries

Marijn Speelman and Ben Kröse
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
contact: b.j.a.krose@uva.nl

Abstract

This paper presents a study on a natural interface game in the context of a library. We developed a camera-based Virtual Mirror (VM) game, in which the player can see himself on the screen as if he looks at a mirror image. We present an overview of the different aspects of VM games and technologies that can be used. For this study a novel framework and prototype game VMQuiz was developed. The prototype was used to evaluate the difference in experienced enjoyment between gaming via our camera interface, and a traditional mouse based interface. The results from the evaluation show that children like playing games in the library and like book related games. The VM game was not experienced to be more enjoyable than the mouse version.

1. Introduction

Over the past years, gaming has become one of the main activities of children and young adults. Using desktop PCs, consoles like the Playstation and Nintendo Wii, and more recently mobile phones, kids spend a great deal of time playing games, alone, over the Internet or with friends in real life. A study done by the NICAM[1] in 2003 shows that almost all kids play a game once in a while and two-third play games on a daily basis. With our life increasingly taking place in the digital world, the popularity of reading books and consequently library visits by children is decreasing.

Games and books have many similarities that make games interesting for libraries. Games can be informative, or entertaining, like books. We were interested whether playing games inside a library would encourage children to visit the library more often, and keep them in contact with the 'old' medium.

Playing games in (semi)-public places puts some constraints on the interfacing. Traditionally gaming is done using a gaming console using some sort of joystick or a desktop PC using a mouse and keyboard. New ways of interaction emerged over the years and have reached the consumer market in great numbers. The Nintendo Wii [11] is the latest example of a successful product that has a non-traditional interface. The Wii uses sensors to detect the orientation of the controller. Control can also be done using computer vision techniques like motion detection. Such games can be called webcam or camera-based games. If the player sees himself on the screen, it is called a Virtual Mirror (VM) game [31]. Sony has developed the EyeToy [7], which is a small USB webcam that can be attached to the Playstation 2 to play these kinds of games.

In particular webcam games could be an interesting addition to libraries. They are distinctly different from traditional games because they are relatively new and unknown, provide a new way of interaction and are not available for every child in their home.

Following these thoughts, the goal of our research is: 1) To compare the level of enjoyment between traditional gaming and VM gaming and 2) to determine if the library is a suitable environment to deploy such games. To test these hypothesis we created a prototype game named *VMQuiz*. A software framework was developed for this game and to use in future studies on webcam gaming. It supports rapid development of multimedia applications and implementation of existing computer vision software libraries.

In Section 2 we describe how gaming fits the context of a library. In Section 3 we describe how computer vision can be used in gaming and which technologies and approaches of movement analysis can be used. In Section 4 we describe the technical and functional aspects of the prototype that has been created. Finally, Section 5 presents the experiment we conducted to evaluate the prototype and the results of these experiments and Section 6 the conclusion.

2. Gaming in Libraries

Libraries are not educational institutions like schools, but they do have a great value in teaching children certain values, skills and increasing general knowledge. One of the main goals for libraries is of course to increase literacy for

[1] Nederlands Instituut voor de Classificatie van Audiovisuele Media

children. Games can be a great addition to books. In a certain way they are 'just another' medium next to books. One of the most clear similarities is that both books and games often try to tell a story. The writer tries to let the reader feel as if he takes really takes part of what is happening in the story. Story lines and genres in books and games are often closely related. For example there is the game Warcraft by Blizzard which is inspired by The Lord of the Rings by J.R.R Tolkien. The story line of Warcraft has even been expanded in a book series. This shows games and books can be closely related. Apart from their links in genre and story telling, playing games could increase skills like reaction time, resource management, team work, critical thinking, making quick decisions, fast reading and coping with stress. It can also increase knowledge on subjects like historical times [26].

Libraries acknowledge the value of games and have been lending out games to customers for many years. The last few years libraries have been more actively using games inside the library. An overview of the intersection between gaming and libraries is given in [23]. Many libraries have PCs for kids with pre-installed games. Often these games have some educational value, like teaching basic math skills using a fun story and interesting setting that relates to the mind-set of children. Next to their educational purpose, libraries hope games make a stay at the library more fun. Although libraries maybe do not have that 'dusty' reputation it had in the past, there is still much work to be done. There are large groups of children that do not visit the library at all.

In the United States there have been several libraries that held LAN parties on weekends or that have been trying to put innovative gaming applications on the library floor. In the Netherlands there have also been libraries, like the library of Vlissingen in the province of Zeeland, that organised gaming days [9]. There are several websites in the form of communities with blogs and wikis where librarians share their experience on gaming activities like LAN parties[2] [1]. An important aspect of these gaming events is the social interaction between the players. This effect can be even stronger with games with a physical interface, like the Wii or Virtual Mirror games.

3. Computer Vision in Gaming

A new way of controlling a game is to use camera input (often a webcam) and computer vision techniques. The computer program tries to understand the contents of a scene by determining specific features like faces or by detecting motion in a stream of images. The result of this analysis can be used to create a game that is based on physical movement. In the following sections we describe such a system and what kind of techniques for analysis can be used.

[2]A Local Area Network (LAN) party is a gathering where several gamers play games against each other

3.1. Virtual Mirror Effect

A Virtual Mirror (VM) game is a type of webcam game where the player is being recorded using a digital camera and sees himself in real time on the screen or projected image in front of him. This creates the illusion of looking at a mirror. The result is a combination of the camera image and (virtual) game graphics. A graphical representation of a physical setup of such a system can be seen in figure 1.

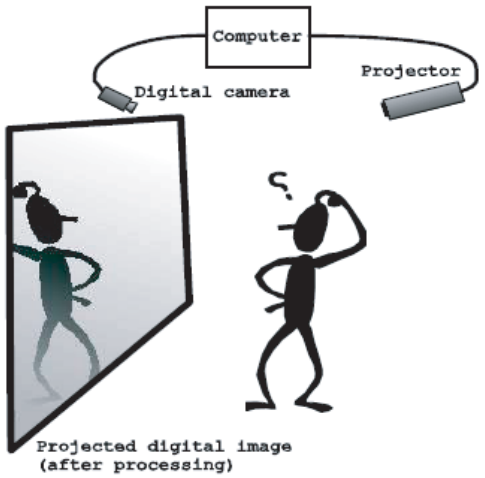


Figure 1. Virtual Mirror setup, from [31].

One of the first descriptions and applications of a VM game was VIDEOPLACE (1985) [21]. VIDEOPLACE was designed to be an addition to traditional telephony so both players could see each others silhouette, but it was also used to play a little game. A creature called CRITTER climbs the silhouette of the player, slides down on his arm and repeats this.

The mirror effect comes with some difficulties. Because the image is recorded from a static location, the eyes of the user do not follow the image as with a real mirror [31]. A problem related to this is eye contact. With a real mirror the person keeps eye contact as he moves in front of the mirror. There are some technical and practical solutions for this, but those are not really needed for intuitive interaction. Especially when playing a game the player will not really notice where his eyes are looking at because he is most likely looking at his hands or game graphics on the screen.

The game graphics can be virtual objects that can be manipulated. The objects are computer generated and react to the movements of the player. There are different types of objects:

- Static objects. This is the static game interface of the

game, like the current score of the game.

- Buttons. These objects can be activated or touched by a player. By generating activity by moving at the location of the button, it is being pushed.
- Movable objects. These objects react to the movement of the player. By analysing the players movement the player can drag the images across the screen.

In the next section we will discuss how interaction with the aforementioned buttons and movable objects can be realised.

3.2. Technologies for VM Games

There are several techniques that can be used to detect and analyse player movement and player features. In this section we describe the techniques that are most relevant for VM games. First, there is simple motion detection that uses image differencing to determine moving areas. Secondly, there is face detection to detect a human face and thus the presence of a player. Finally, there is object detection, which is the detection and tracking of predefined objects using shape or colour.

Other frequently used techniques are background subtraction and skin colour segmentation, which are beyond the scope of this paper.

3.2.1 Motion detection

Motion detection or motion sensing uses simple image differencing. The RGB colours are compared between two subsequent frames. When the difference between the RGB value of two corresponding pixels is greater than a certain threshold, the game detects movement. These single changed pixels can then be grouped and used to determine larger areas of movement. This type of motion detection can be used to touch objects in a game, like hitting a ball with the head or foot (Virtua Soccer [8]) or touching a virtual button. Because the game does not know the direction of the motion with this method, it can not be used to push around movable objects.

To track the direction and speed of motion, optical flow can be used [13]. For each frame the displacement compared to the previous frame is calculated for certain areas. These areas can be the virtual objects that are being tracked. It is not needed to calculate the optical flow for the whole image. A comparison between motion detection and optical flow is done in [31]. The study shows using optical flow for touching virtual buttons performance faster and more accurate than motion detection, especially for complex menus.

Often when analysing body movement with a camera there is no distinction made between different limbs like arms or legs, just the relative frame by frame difference is used. One can look at the movement of the limbs from the front of the player, or at the movement of the whole body in space from the top. An example of a combination of both is the experiment in [22] where a wide-angle lens looking down onto the playing field was used to keep track of spatial orientation in the map of the game. Also the movement of the arms was used to execute different game commands.

3.2.2 Face detection and -tracking

Face detection should not be confused with face recognition. *Detection* only detects if and where a face is present in the image, while *recognition* tries to determine who's face it is. Face tracking is done by detecting the face in each subsequent frame.

There are various methods for the detection of faces in images or sequences of images. An extensive survey of current face detection methods is given in [30]. In our prototype game (section 5.1), we used the face detector proposed by [28] to detect frontal faces. It is scale-invariant, meaning all sizes of faces can be detected in the image.

The added value of face detection in gaming was studied by [29]. It concluded that the use of face detection has several advantages compared to traditional motion detection. The main two reasons are:

- Increase of the ability of the game to detect if a person is present and the location of that person.
- Improvement of the game experience using role playing by adding virtual objects on the player's body, for instance a hat or jewels around a player's neck.

Other advantages are that the player does not have to move to let the game detect its presence, the player can be recognised without doing anything. This could be used to automatically start the game when a player enters the game area. With face detection it is also possible to detect the number of separate players so each player can be given a separate role in the game. Disadvantages are that face detection is more reliant on external factors like skin colour and lighting. The calculation time is longer and implementation is more difficult than when using motion detection.

Face tracking and object tracking are closely related. A face is actually just an object with certain detectable features, like skin colour and the area of the eyebrows, eyes, nose and mouth. Faces are more dynamic than most objects because they can change, for instance when someone laughs.

3.2.3 Object tracking

Real objects can be used to interact with the virtual world of a game. An example is using a real paint brush to paint a virtual painting [6]. Giving the object a distinct colour

(like bright green) makes it easy to locate and track. Object tracking is less dependent on the background than motion detection, but can be more sensitive to changing light conditions. A comparison between motion detection and object tracking is done in [17]. The study concludes that, especially using many virtual buttons, object tracking performs faster and more accurate. A disadvantage of using objects is the need to constantly hold the object. Another possible reason to not use objects, is that the player only has to move his arms and not his whole body.

3.3. Existing webcam & Virtual Mirror games

There are several existing webcam games. The oldest camera-based game is considered to be VIDEOPLACE, discussed in section 3.1. Other applications are the Vivid Group Gesture Xtreme System (1986) [4], the Reality Fusion GameCam (1999) [5], the Sony EyeToy (2004) for the PlayStation2 [7] and Microsoft Xbox Live Vision (2006) [10]. Motion detection and face recognition are the most used computer vision technologies in these games. For example the Xbox Live Vision software supports capturing the face of the player and using the face on a virtual character in the game. Over the past few years, webcam producers like Logitech and Philips have also been implementing face detection and tracking in their webcam software and including small games based on gestures. This shows these technologies are becoming more common for consumers.

There are also small Flash games [2] that can be played instantly inside a browser and only require the Adobe Flash plugin and a webcam. No additional hardware or software is needed, unlike the EyeToy and Xbox Live Vision camera. Most of these games use only primitive image differencing motion detection.

4. Game evaluation

Evaluating games can be a complicated task. For game developers it is increasingly important to evaluate the games during development and get constructive feedback to deliver enjoying and well selling games. Like in this study, game evaluation can also be a research tool to compare certain aspects or features of games.

There are many terms that are related to the experience of playing a game, for example: enjoyment, usability, fun, playability and flow. In more general terms, *fun* is the most important aspect of whether someone likes to play the game or not. If a game is not considered to be fun it is simply not being played, or at least not for long. Another common term is usability, which is defined by the ISO 9241-11 standard as a combination of three measures: effectiveness, efficiency and satisfaction. A brief overview of the link between usability and games is given in [20].

Different tools can be used to evaluate games. Cameras are useful to record the behaviour of the player. The video can later be used to analyse movements and to detect emotions like frustration and confusion. It is also possible to gather in-game statistics. For example the score, playing time and game actions of a player can be used to make assumptions about the difficulty and enjoyment of the game. Interviews can be used to gather qualitative information about the game or get statistical data. Questionnaires can be used to gather quantitative data about the game.

A game can be evaluated using expert analysis or user studies. Heuristics are often used with both. These heuristics are based on theoretical and empirical research. One of the first empirical studies we found on the use of heuristics and game evaluation is [24]. Different versions of a dart game were compared during a user study. The total time playing the game was taken as a indication of the enjoyability. From these results heuristics were extracted. Over the years several other studies on heuristics have been done. See [18] for an extensive overview of usability heuristics literature in gaming. Other studies are [16] in which the measurement of playability with heuristics is discussed, and GameFlow [27].

The GameFlow Model heuristics are derived from literature (i.e. [18] [16]) and is combined with the *flow* theory [15]. The idea of flow is that a person is fully involved in the experience and completely focussed on the task it is doing. When a player has this experience, the enjoyment of a game is likely to be greater and the game will be played longer. A short and clear overview of flow and gaming is given in [14]. In the GameFlow model eight criteria are given to measure the enjoyment of a game: concentration, challenge, skills, control, clear goals, feedback, immersion and social interaction. We based our questionnaire the criteria 'control', 'skills' and 'immersion' because only the criteria that are related to the interface and control were used. Furthermore we added questions to measure enjoyment directly, from a study on the enjoyment of interaction [19] and motivation [25].

The study done in [12] is comparable to our study in the sense that the difference between two game interfaces is researched. They developed an outdoor Augmented Reality (AR) Space Invaders 3D game and compare that with a PC version of the game. A questionnaire with 19 questions and likert-scale answers is used to measure the satisfaction. The AR version appears to be more satisfying than the PC version. Where comparing two versions, a between-subjects or a within-subjects design can be used. This means either of the versions can be played by each participant or each participant can play both versions. The advantage of a between-subjects design is that the order of the played versions and previous version does not influence the result. We used a between-subjects design because of this. The disadvantage is that it is not possible to compare the opinion of

each player for both versions.

5. VMQuiz, a prototype VM game

A VM prototype game was created to compare the experience with a traditional mouse interface and a physical interface that uses the webcam. The game can be played with either of the input devices. A framework using a combination of Adobe Flash and C++ was developed to be able to quickly create games that include animations, sound and vector graphics, while at the same time maintaining the opportunity to use advanced image analysis.

The game was developed in cooperation with the Bibliotheek van Almere, the main library of the city Almere, Netherlands. Beforehand a list of important aspects and requirements of the game was put together. These requirements originated from meetings with the library, existing game heuristics from literature and personal gaming experience.

In the next section (5.1), the game itself is described, as well as some design choices. In section 5.2 the technical implementation of the game is discussed. Finally, some library specific aspects are discussed in section 5.3.

5.1. Description of the game

We have chosen to create a book quiz that asks questions about popular books for children. Children are familiar with quizzes and the goal of the game can be easily understood: giving the correct answers to the questions. Questions were related to well known books like Harry Potter, Pippi Langkous, De Kameleon, and (Dutch) book authors like Paul Biegel and Carry Slee.

The game was developed for boys and girls from the age of 7 to 11 and the questions were adjusted to that age group. There are two types of questions, *Multiple Choice (MC) questions* and *Sort questions*. The game consists of 12 MC questions and 3 sort questions. Most of the questions were created by the library of Almere. The number of MC questions is greater than the number of sort questions, because they are faster to answer. The questions were mixed in advance to give some variation. With the MC questions, the player has to pick one of the four possible answers. An answer is chosen by sliding it to the side of the screen, as shown in figure 2 (webcam input). In figure 3 a screenshot is shown of the version that uses mouse input. The Sort questions ask the player to match 4 images and words. The images have to be dragged to the designated drop zone beneath the corresponding words. The images originate from books. An example can be seen in figure 4.

The webcam version requires the player to move across the screen. Often ducking or crawling is needed to give an answer. The player stands in front of the webcam and sees himself on the projected image on the screen. The game

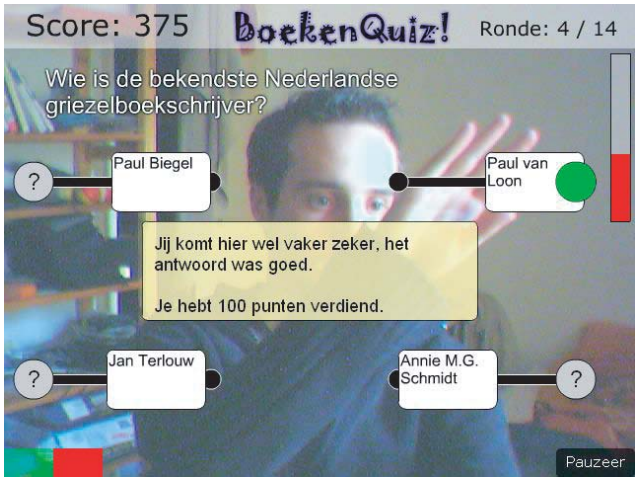


Figure 2. Webcam: Answering a Multiple Choice question by sliding the correct answer to the corner.

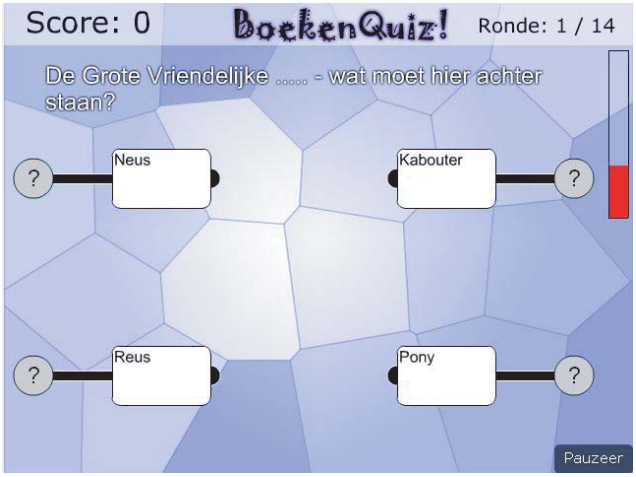


Figure 3. Mouse: Screenshot of a Multiple Choice question played using the mouse.

graphics are added to the webcam image. The mouse version works by just clicking or dragging the objects with the mouse.

The game is aimed to take about 5 to 10 minutes to play. That should be enough to give the child a good idea of what the game is like and at the same time will not put too much stress on the child. For each question a bar is displayed in the top right corner to show the remaining time of the current question. When the time is up, the player can still give the right answer but receives only half of the points he would normally get. For the MC questions the player has two tries to get it right. Giving the correct answer in the first attempt within the time limit gives 100 points, a correct answer the second try gives 75 points. When two wrong answers are given the correct answer is shown and the game continues with the next round. The sort questions give 100 points. When an image matches the word it is hovering at

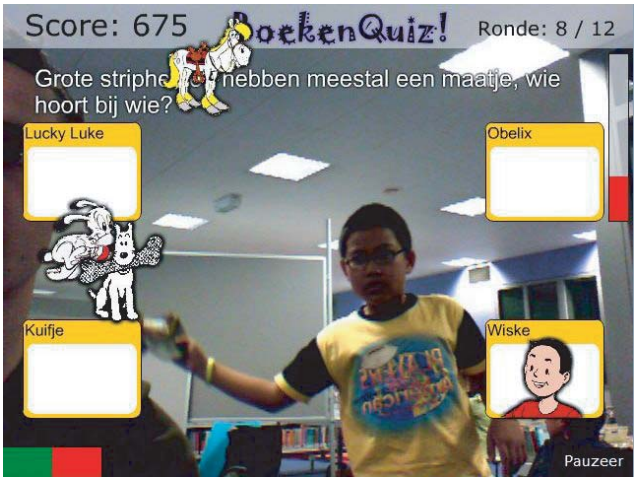


Figure 4. Child answering a sort question by pushing the images (the two dogs, horse and head) across the screen.

that moment, the image locks to the word and can not be moved anymore. For example, the bottom right image in figure 4 is locked to the white coloured dropping zone.

The progress of the game is shown in the top of the screen. The current round and total number of rounds gives an idea of how long the game will be. The current score is given to show the progress of the player and to increase the motivation to get a higher score.

For most questions a small sound sample is played that relates to the current question. Examples are the sound of a howling wolf with horror questions and the Harry Potter tune with a question related to Harry Potter. Giving the wrong or correct answer plays respectively a buzzer or trumpet sound to give clear feedback of the players actions. When all questions are answered the game is finished and the final score is displayed on screen.

5.2. Technical implementation

A framework was created to allow rapid development and expansion of the current prototype. The framework consists of a front-end game client (Adobe ActionScript 3) and a server (C++) that analyses the image stream from the webcam. The analysis is done using the OpenCV computer vision library [3]. See figure 5 for a graphical representation of the application components. To minimise network delay, all applications are run on the same computer.

The client takes care of the game graphics, sounds and actual gameplay. For the client the Adobe ActionScript language was chosen because it is a high level scripting language with support for many multimedia formats and is very suitable for game creation, animation and vector graphics. The latest version, ActionScript3 was used because it is faster and has better functionality for accessing raw image and network data than previous ActionScript

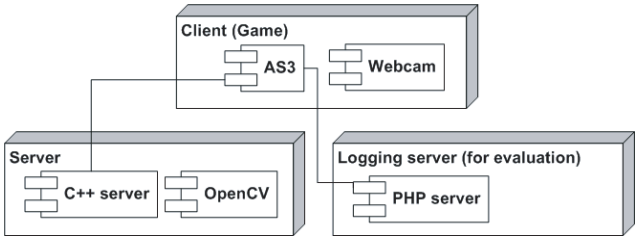


Figure 5. Application components of the framework.

versions. ActionScript code is compiled into a executable Flash file (.swf) that can be run in the browser. The server has to be started separately. Additionally a simple PHP[3]-script was used to log the game progress and scores to give insight in the amount of time that is needed to complete certain questions and the game as a whole.

The webcam input is read by the Flash client and sent to the server for analysis. To have smooth movement feedback, it is necessary to send image data very frequently. For face detection the image data is sent every 50ms. For the optical flow calculation the data is sent every 30ms. On a computer with a 2.8Ghz Pentium 4 processor, the average time needed for face detection is 20ms. The optical flow takes about 8ms to calculate. On top of this, there is an overhead of 10-15ms for transferring the content between the applications.

Each frame is sent in a package that contains the following data (total of 77609 bytes):

Mode A number representing the mode of the game, which can be *detect face*, *optical flow* or *inactive*.

Image Data A scaled down bitmap of the current frame of the webcam stream. The original webcam size is 640x480 pixels which is re-sized to 160x120 before sending to speed up both the sending of the data as well as the analysis of the image.

Number of Objects to Track A number which tell how many virtual objects there need to be tracked.

Object locations The x and y coordinates of the virtual objects being tracked.

Just a black and white image is needed for face detection and optical flow, so the colour channels are not used. We used the face detector that is available in from the OpenCV library, as discussed earlier in section 3.2.2. The face detection is used to automatically start the game when a players' face enters the playing area. The movement of objects is done using optical flow as discussed in section 3.2.1. After analysis the result is sent back to the client. For face detection the x and y coordinates, width and height of the

[3]Website: http://www.php.net

detected face are transmitted. For optical flow the *new* locations of the virtual tracked objects are sent back to the game client and processed further.

The mouse game is the exact same as the webcam game, with the single difference that the mouse game has the image analysis turned off and instead shows a background image and not the webcam stream.

5.3. Playing the game inside a library

When playing a game in an environment like a library, there are things that have to be taken into account:

- **Starting the game** The game should be played with as little mouse or keyboard activity as possible. For instance it should be able to start automatically when a player enters the game area. This can be done using face detection or simple motion detection. Face detection is more accurate because then a passing person will not trigger the game by accident.
- **Background** The game should be set up in a way that background movement is minimised. A room divider can be used to shield off the game area. Background subtraction and face or body recognition could be used to filter out background noise. The background colours, background pattern and lighting conditions can greatly influence the performance of the image analysis.
- **Position and distance of the camera** Different setups are possible. The player must be able to reach the corners of the screen and must be able to duck or crawl while staying in sight of the camera. The height of the player also affects the ideal distance and panning of the camera.
- **Movement space** The player should have plenty of room to move around, without accidentally touching objects or other people.
- **Sound** When playing a traditional game it is easy to use a headphone or small sound boxes so other library visitors will not be interrupted. Children playing a webcam game are further away from the screen and computer which could result in a more noisy environment.

An example setup can be seen in figure 6. Using a curved wall the sound will be somewhat contained in the gaming area and there will be minimal background noise of spectators or people walking by. The cross is where the player stands, watching to the screen in front of him where the image is projected on. Aiming the webcam at eye level will improve the mirror effect. The beamer can be attached to the ceiling or placed in front of the player. The webcam can be positioned on different locations:

A. Beamer and camera on the table in front of the player.

B. Camera to the left or right of the projected image.

C. Camera recording at eye level through a little hole in the screen.

D. Camera behind a semi-transparent screen.

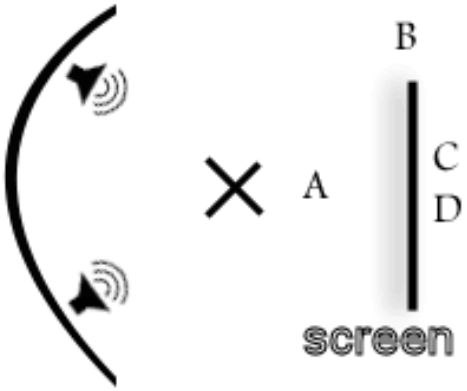


Figure 6. Example of a webcam game setup.

6. User study

The goal of our user study is to determine the level of enjoyment for both versions of the game. The second goal is to gain experience with these kind of webcam games and to find out if children enjoy playing games that relate to the books they read.

Our study researches the following hypotheses:

1. The webcam interface version is experienced to give more enjoyment than the traditional mouse interface version.
2. Children like it when they play a game that relates to the library context they are playing in.

6.1. Participants

The experiment was executed during two days at the Bibliotheek Almere, the Netherlands in December 2007. In a relatively quiet corner of the library we set up a beamer, projection screen and laptop with integrated webcam. To minimise background noise the area was partially excluded by a room divider, as seen in the background of figure 4. In figure 7 a picture of the evaluation setup is shown.

We recruited 60 children as participants of this study. Half of them, 30 children, played the webcam version; the other half played the mouse version. This results in a between subjects setup. One-by-one the participants were put in into the two different categories. Recruitment was done

Figure 7. Child evaluating the game inside the Bibliotheek Almere.

using convenience sampling in the library and an announcement on the website of the library. This announcement was also copied by the local newspaper.

The group of participants consisted of 27 boys and 33 girls. The mean age was 9,9 with a standard deviation of 1,7. The participants did not receive a reward for their participation.

6.2. Procedure

Each participant was asked to play one of the versions of the game. First a short explanation was given. Then it was explained the game was about well known books. After that the children were told how to control the game. Depending on the height of the player the camera angle was adjusted and the player positioned in the centre of the screen.

Then the game was played until it was completed. After the completion of the game the participant was asked to answer a few questions about the game. When there was unclear behaviour of the game or the participant became visibly frustrated, additional instructions were given. The mouse version took around 5 minutes, the Webcam version took 5 to 10 minutes to play. The answering of the questions took about 5 minutes for each participant.

6.3. Instruments

After playing the game, the participants were interviewed. Based on the literature study in section 4 we constructed a questionnaire. To measure the enjoyment we used relevant concepts from the GameFlow heuristics, namely Control, Immersion, Skills, with the general concept Enjoyment. The result is the list of 15 statements as showed in Table 1. To test the second hypothesis, 3 questions that are related to the content and context of the game, respectively books and the library are added. Answers were based on a 3 point likert-scale. The questions were posed in random order. Both versions recorded the game progression, like the time of the whole game and each separate question. The game scores of the player were written on the questionnaire form.

Table 1. The questionnaire used in the experiment

Control	
C1	I found it difficult to control the game. *
C2	I found it easy to pick the right answer.
C3	I understood quickly how to control.
Immersion	
I1	Some moments I was really concentrated on the task.
I2	I had the idea that I was completely into the game.
I3	Some moments I forgot that I was in the library
Skills	
S1	I played with a Ninetendo Wii before.
S2	I play computer games.
Enjoyment	
E1	I enjoy playing the game.
E2	I find the game enjoyable.
E3	I find the game fascinating.
E4	I find the game boring. *
E5	I am glad that the game is over. *
E6	I would like to come back and play the game again.
E7	I experienced the game as schoolwork. *
Library specific	
L1	Did you know the books the questions were referring to?
L2	Did you like the fact the questions were related to books?
L3	Would you visit the library more often when you could play this game and other similar games?

* Contra indicative

After these questions, the children were asked three open questions about what they liked and did not like about the game, and if they knew of something that could improve the game. Finally the background variables age and sex were collected.

6.4. Results

All children in the test completed the game. Some children had difficulties to complete the sorting part of the quiz. With some help of bystanders they were able to continue to the next round. Overall the group of children was quite experienced with playing computer games. More than half of the participants had experience with the Wii and played games more than 4 times per week. The average score of the players was 760 with a standard deviation of 200.

We executed a reliability analysis on the groups of question to assess if they measure the same concept. For this the

Table 2. Mann-Whitney U test comparing the two different interfaces

Code	Mean webcam	Mean mouse	Mann-Whitney U	Asymp. Sig. (2-tailed)
C1	2.07	1.17	182.500	.000*
C2	1.87	1.50	320.000	.034*
C3	1.50	1.13	324.000	.016*
I1	1.23	1.23	438.500	.812
I2	1.53	1.87	327.000	.047*
I3	2.40	2.07	347.500	.097
E1	1.07	1.00	420.000	.154
E2	1.13	1.07	434.000	.621
E3	1.47	1.27	404.000	.397
E4	1.03	1.00	435.000	.317
E5	1.53	1.43	432.000	.759
E6	1.37	1.33	415.000	.504
E7	1.47	1.67	380.000	.241
L1	1.80	1.90	412.500	.482
L2	1.20	1.30	406.500	.369
L3	1.30	1.23	445.000	.926

* Significant at the 0.05 level (2-tailed).

Cronbach's alpha statistic was used. For this test we used the acceptable value of an alpha higher than 0.7. The Enjoyment concept had a Cronbach's alpha of .719 after deletion of question E7. None of the other concepts showed a sufficiently high alpha, which means no questions could be grouped together.

Correlations between questions and variables were calculated using Spearman's rho. What clearly stands out is that there is a strong negative correlation (ρ = -0.773, p = 0.000) between the played version and C1. This means the webcam interface was experienced to be far more difficult to control than the traditional mouse interface. We tested this with a nonparametric Mann Whitney test, the results are shown in table 2

For research hypothesis 1 a nonparametric Mann Whitney test was done with the two 30 participant groups of players. Question E1 to E7 were used for this. We also studied the Immersion concept and the library related questions. A score of 1 means 'agree' for positive questions and 'not agree' for negative questions. The result is shown in table 2. For the Control concept the results correspond to the results of the Spearman's test. It is significant that the webcam version was found more difficult to control. The mean score for the enjoyment questions was slightly higher in the webcam condition than in the mouse condition, but the difference is not significant.

For research hypothesis 2 the mean score was calculated and compared from the last three questions (L1, L2, L3). As shown in table 2 there was no significant difference between the two interfaces. The children really like the fact that the questions were related to books. They also believe they will visit the library more often if they could play such games. Not surprisingly, there is a positive correlation between age and L1, the familiarity with the books the questions were referring to. In general these three questions show the children have a positive attitude towards library or book related games.

6.5. Discussion

The goal of this experiment was to measure the enjoyment of both interfaces. A higher score on the questionnaire should result in a better and more enjoyable game. In general there were no significant differences between the two interfaces. This suggests that both interfaces have the same perceived enjoyment. Another explanation is that the questionnaire is not entirely suited for correctly measuring enjoyment and flow. It is likely that there are not enough questions to get an accurate measurement of all aspects. The results of the questions appear to be highly influenced by the experienced difficulty to control the game.

There were several children that played both versions of the game. The interview was done before playing the other version to not influence the results. All children that played both versions said they liked the webcam interface more, because they liked the movement aspect of it. They were visibly enthusiastic when playing the game, with a few exceptions. One participant got quite frustrated with a sort question and almost gave up. He told us he was used to the capabilities of the EyeToy, which allows faster and less accurate movements.

The questionnaire results show that the children liked the idea of playing book related games. This also shows from the collected comments. Many children said they enjoyed the game because it asked questions about books they know. Some children also said they learnt some new facts about books. They even expected to visit the library more often when these kinds of games can be played. They were also interested in playing our game again.

The played game was a prototype that had not been played by children before. It had some short-comings that negatively influenced the enjoyment of the game. For example the controls for the sorting questions were found too difficult for a number of children. Further improvement and a more robust movement analysis is needed to present a fair 'opponent' to the more easily controlled mouse interface. Looking at the scores and comments of the children, the difficulty of the game questions seems to be right. The age of the child shows a correlation between how much the children know about the books, but this did not affect their score enough to be significant.

7. Conclusion & Future work

In this paper we presented an overview of the current state of webcam games and the techniques that can be used to analyse the webcam images. A literature study was done on the evaluation of games, which showed there are multiple studies that use heuristics for evaluation but none really fit webcam games well. We showed how gaming and in particular webcam games can be played in libraries and what the possible benefits are, like more frequent library visits and making visits more fun for children.

Based on the results of the user study using our VMQuiz prototype game, we can conclude that there was no significant difference between the measured enjoyment of the player playing the traditional mouse game, and the one with the webcam interface. Both versions scored high. After the experiment children were allowed to play both conditions. From the comments we collected during the interviews, the children appeared to enjoy playing webcam games and thought it was a fun way to interact with the game.

The second research objective was to test the hypothesis that children like playing games that relate to books and that they like playing them in the library. From the results of the questionnaire it can be concluded that this hypothesis is valid. The children were also visibly enthusiastic when playing both versions of the game and predicted they will visit the library more often when these kind of games can be played in the library.

Finally, a novel software framework was developed that aims to make development of webcam games easier and quicker and make use of more advanced computer vision algorithms. It consists of a Flash client for the actual game and media like sound and graphics, and a C++ server application that is able to use existing image analysis algorithms.

The research has raised a number of issues with the current prototype and questionnaire. To further analyse the effect of the game interface, the prototype should be improved. Movement analysis can be made more robust by improving the current methods and using more accurate movement tracking algorithms. Also different types of questions and game objectives can make the game more easy and fun to play. The questionnaire can be expanded with more questions about the specific interface. Questions that ask more directly about the enjoyment could also be added.

8. Acknowledgements

Thanks to Zoran Zivkovic and Francesca Hagethorn of the University of Amsterdam for their help and support. Thanks also to Anique Persoon who spent her Christmas holidays in the library doing a full set of experiments. We thank Soan Lan Ie and Peter Nugteren from the Bibliotheek Almere for their time, effort and enthusiasm they put into this project. Finally we would like to thank the Waag Society[4] for sharing their experience on game development. thank Alex de Vries for his C++ code that contributed to this project.

References

[1] Libgaming google group.
`http://groups.google.com/group/LibGaming/`

[2] Newgrounds.com, webcam games.
`www.newgrounds.com/game/webcamgames`

[3] Opencv software library.
`www.opencv.org`

[4] Vivid group inc.: Gesture xtreme system, 1986.
`www.vividgroup.com`

[5] Reality fusion inc: Gamecam, 1999.
`www.vividgroup.com`

[6] Setpixel.com: Magic coloring wall, 2002.
`www.setpixel.com/content/?ID=twothingsdotone`

[7] Sony computer entertainment inc.: Sony eye toy, 2003.
`www.eyetoy.com`

[8] Sega superstars, eyetoy games, 2004.
`www.sega.com/gamesite/segasuperstars/`

[9] Gaming lan-party in the library of vlissingen, netherlands, 2006.
`www.vlissingen.nl/web/show/id=291737/contentid=12881`

[10] Microsoft xbox live vision camera, 2006.
`www.xbox.com/en-US/hardware/x/xboxlivevision/`

[11] Nintendo wii, 2006.
`http://wii.com`

[12] B. Avery, W. Piekarski, J. Warren, and B. H. Thomas. Evaluation of user satisfaction and learnability for outdoor augmented reality gaming. In *AUIC '06: Proceedings of the 7th Australasian User interface conference*, pages 17–24, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc.

[13] S. Beauchemin and J. Barron. The computation of optical flow. *ACM Computing Surveys*, 27(3):433–467, 1996.

[4] Website: http://www.waag.org

[14] J. Chen. Flow in games (and everything else). *Communications of the ACM*, 50(4):31–34, 2007.

[15] M. Csikszentmihalyi. *Flow: The Psychology of optimal experience*. Harper and Row, 1990.

[16] H. Desurvire, M. Caplan, and J. A. Toth. Using heuristics to evaluate the playability of games. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 1509–1512, New York, NY, USA, 2004. ACM Press.

[17] J. Eisenstein and W. E. Mackay. Interacting with communication appliances: an evaluation of two computer vision-based selection techniques. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1111–1114, New York, NY, USA, 2006. ACM Press.

[18] M. A. Federoff. Heuristics and usability guidelines for the creation and evaluation of fun in video games. Master's thesis, University Graduate School of Indiana University, 2002.

[19] M. Heerink, B. Kröse, B. Wielinga, and V. Evers. Enjoyment, intention to use and actual use of a conversational robot by elderly people. In *Proceedings of the third ACM/IEEE International Conference on Human-Robot Interaction*, Amsterdam, 2008. ACM.

[20] A. H. Jorgensen. Marrying HCI/Usability and Computer Games: A Preliminary Look. In *Proceedings NordCHI'04*, pages 393–396, Tampere, Finland, Oct. 2004. ACM.

[21] M. W. Krueger, T. Gionfriddo, and K. Hinrichsen. Videoplace an artificial reality. In *CHI '85: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 35–40, New York, NY, USA, 1985. ACM Press.

[22] S. Laakso and M. Laakso. Design of a body-driven multiplayer game system. *Comput. Entertain.*, 4(4):7, 2006.

[23] J. Levine. Gaming & libraries: Intersection of services. *Library Technology Reports*, 42(5), 2006.

[24] T. W. Malone. Heuristics for Designing Enjoyable User Interfaces: Lessons from Computer Games. In *Proc. 1982 Conference on Human Factors in Computing Systems*, pages 63–68, New York, 1982. ACM.

[25] S. Reeve. The interest-enjoyment distinction in intrinsic motivation. *Motivation and Emotion*, 13(2):83–103, 1989.

[26] K. Squire and H. Jenkins. Harnessing the power of games in education. *Insight*, 3(5):2–30, 2003.

[27] P. Sweetser and P. Wyeth. Gameflow: a model for evaluating player enjoyment in games. *Comput. Entertain.*, 3(3):3–3, 2005.

[28] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2002.

[29] S. Wang, X. Xiong, Y. Xu, C. Wang, W. Zhang, X. Dai, and D. Zhang. Face-tracking as an augmented input in video games: enhancing presence, role-playing and control. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1097–1106, New York, NY, USA, 2006. ACM Press.

[30] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *Pattern Analysis and Machine Intelligence, IEEE*, 24(1):34–58, January 2002.

[31] Z. Zivkovic. Optical-flow-driven gadgets for gaming user interface. In *Entertainment Computing ICEC 2004*, pages 90–100, 2004.

An Online Face Avatar under Natural Head Movement

Haibo Wang[†,‡], Chunhong Pan[†],Christophe Chaillou[‡], Jeremy Ringard[‡]
[†]LIAMA & NLPR, Institute of Automation, Chinese Academy of Sciences
[‡]LIFL & Inria Futurs,University of Lille, FRANCE
{hbwang1427,chunhongp}@gmail.com,{christophe.chaillou,jeremy.ringard}@lifl.fr

Abstract

Creating a realistic face avatar is still a challenging problem. In the paper, we propose a new video-based technique for synthesizing such an online avatar that is capable of replicating facial expressions under natural head movements. Our approach is to track 3D head pose, simultaneously extract video face textures from monocular video sequences and then map them onto a static head model. In contrast to most of the current facial animation solutions, our approach can avoid the procedures to track high dimensional facial features, warp face textures and deform face model. In addition, the method allows independent control of head pose and facial textures that are separated from videos. The demonstrations of our method in real video scenarios validate its efficiency.

1. Introduction

Creating a realistic face avatar is as of today still a great challenge as our computer is deficient to simulate the complicated mechanism of facial behaviors and we are all expert to distinguish even subtly unnatural facial expressions. Especially, for both gaming and collaborative work applications, such as the Second Life or 3D Poker, to create face avatar could be more challenging. To increase users immersive experience, face avatar embodied in those online games should replicate users facial expressions in real- time. In the paper, such a face avatar that needs automatic, realistic and real-time facial expression cloning is named as online face avatar.

The natural way to create an online avatar is from video sequence. Peter Quax [14] simply used raw video stream to replace avatar head in gaming, which cant overcome the dimensionality gaps between 2D video and 3D simulated environment. Recently, extracting only facial expressions from videos are thereby more promising as vision techniques have made progress. The core vision task is how to extract natural facial expressions from the facial actions as well as head movements of the user in video stream. Recent work is to use Active Appearance Model(AAM) [5] or 3D Morphable Model(3DMM) [2] to simultaneously track facial action and head gesture. Then the extracted motion parameters are used to re-animate a dense geometry. The limitations of this method include: the training of tracking models, i.e. AAM, needs tedious manual labeling of many landmarks; it needs to define semantic information of facial features which is still tough; very often tracked facial action and head pose cant be distinguished. In addition, the computational time of non-rigid tracking and geometry animation is often very expensive.

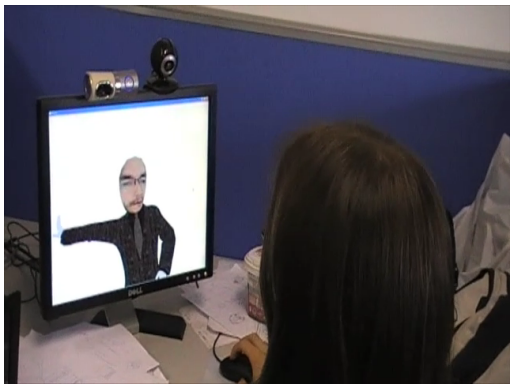


Figure 1. The snapshot of an online face avatar

In the paper, we propose a more efficient scheme to create online face avatar with realistic facial expressions under natural head movements. Our scheme is to extract dynamic face textures from video stream, and then directly map them onto a static head model. The essential idea behind our approach is that: a seamless mapping of dynamic textures onto a static model will come into same expressional realism as dynamic model does. Whereas, direct video mapping enables us to avoid tracking non-rigid facial actions and defining high-level semantic information of tracking parameters. Using static head model doesnt require the time-consuming geometry deformations any more. In addition, as our approach automatically track 3D head gesture, it allows the

user to control the head gestures of his avatar. A similar mapping idea may be found in [18], which addresses 2D facial expression transferring among several identities.

While the above features represent the contributions on the application side, the technical contributions of our scheme are as follows: (1) To ensure online avatar, we modify the Online Appearance Model(OAM) [8] to track 3D head poses. This approach, in comparison with other learning-based methods, doesnt need any training process and can perform the tracking on the fly. We then develop a tracking system with automatic initialization, online pose tracking and tracking recovery; (2) The static head model should be individualized. We develop an easy-to-use graphical interface to support this model individualization by giving two photos.

2. Related Work

In this study, we list the recent trends concerning head pose tracking and facial expression synthesis techniques.

A large body of literature exists for 3D head pose tracking. Previously, motion-based [15, 20] tracking algorithm dominated the field. More recently, combining appearance and shape tracking methods are incrementally grouping up [11]. This line of methods is in fact a modified model-based method as they must construct visual models to describe appearance and shape. And they indeed achieve better tracking results than all the previous ones. The appearance and shapes can be learned either in off-line [11] or on-line [6].

As for realistic facial animation, graphics-based techniques already enables us to create amazingly realistic facial animations. Recently, the realistic facial animations are produced by physically-based [9, 17], motion-captured [1] or laser-scanned [2], which has been successfully applied in game, film, arts and entertainment, as surveyed in [7]. However, graphics-based animation is presently not available for online games as the specific reasons are: physically-based animation is relatively computationally expensive while both motion-captured and laser-scanned animation need expensive performance recording equipments.

Video-based methods provide an alternative to graphics-based animation. One approach is image-based synthesis that can produce animated sequences with a high degree of both static and dynamic realism, by using control parameters to warp static images [16]. Another one is to track facial features from video stream and them use them to animate both the face geometry and textures [13, 19]. More recently, hybrid animation techniques are exceedingly popular. [4] combines video-tracked and motion-captured data to create a rather realistic and real-time face avatar.

3. Scheme Overview

As shown in Figure 2, the proposed scheme is composed of three steps: (1) The left top steps are to online track 3D head pose and extract view-independent face textures. At first, our developed pose tracker automatically tracks head pose from video stream depicting a moving head. Simultaneously, the tracker creates view-independent face textures by warping the covered face regions in video stream; (2) The right top part of diagram is to adapt a generic head model to individuals given his two near orthogonal photos. This interactive process is just done once ahead of online pose tracking; (3) Once finished the above steps, all the extracted video textures are then mapped onto the head model frame by frame to create the online face avatar.

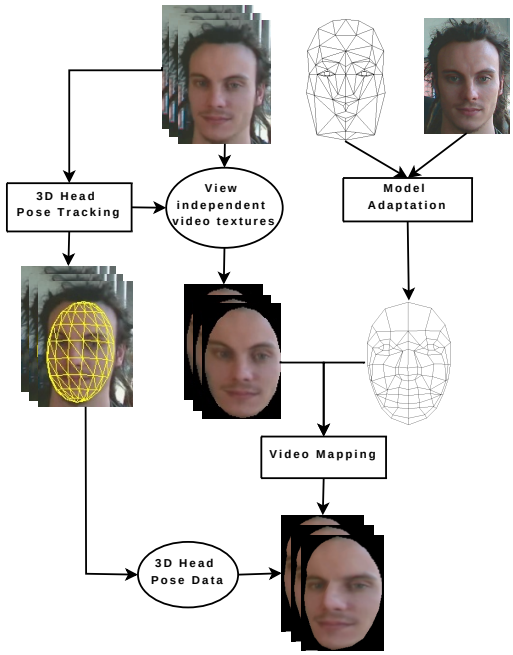


Figure 2. The diagram of creating online face avatar under natural head movements

4. Head Pose Tracking and View-Independent Video Textures

4.1. Online Head Pose Tracking

Track the 6 DOFs of head motion (yaw, pitch, roll and 3D position) is the core element of our scheme. We use a generic ellipsoidal model to approximate the head geometry of the user and then apply an adaptive appearance-based tracking technique to recover his head pose in monocular video stream. This tracking framework follows the algorithm proposed by F. Dornaika [6]. In order to use it for online application, we first propose an pseudo-auto tracking initialization solution. And then we complete the al-

gorithm by adding the strategy of drifting recovery and re-initialization.

Denote the motion vector of head ellipsoidal as $\mathbf{B} = [\mathbf{x}, \mathbf{y}, \mathbf{z}, \alpha, \beta, \gamma]$. Given a video sequence $\{\mathbf{Y_0}, \mathbf{Y_1}, ..., \mathbf{Y_n}\}$ depicting a moving face, the objective of pose tracker is to recover ellipsoidal's motion vectors $\{\mathbf{B_0}, \mathbf{B_1}, ..., \mathbf{B_n}\}$. To estimate current pose $\mathbf{B_t}$, all the historical images $\{\mathbf{Y_0}, \mathbf{Y_1}, ..., \mathbf{Y_{t-1}}\}$ are used to provide estimated information as more as possible. In the tracking context, all the model parameters associated with the current frame will be handed over to the next frame.

4.1.1 Tracking Initialization

Head pose tracking have to initialize the pose and the parameters of the ellipsoidal model, including its radius, height and center position. To our best knowledge, automatic initialization is extremely difficult to set up. Instead, we propose a pseudo-auto initialization that needs an offline interactive operation and can result in an automatic online initialization.

The interactive operation is to fit the ellipsoidal model to two view faces in the beginning of video stream. At the same time, some feature points detected as SIFT keypoints [10] are learned. And then, the online initialization will automatically align the interactively fitted ellipsoidal model by matching the learned SIFT feature points. In the context, automatic model fitting is transferred to a feature matching problem, which is in fact a pseudo-auto procedure. By default, we must ensure that each user keeps a frontal-parallel head pose while the online initialization begins.

4.1.2 Adaptive Appearance-based Head Pose Tracking

The head pose is recovered by minimizing a discrepancy function between an adaptive appearance $\mathbf{A_t}$ and a shape-free texture $\mathbf{X_t}$ warped from the current incoming image $\mathbf{Y_t}$. The adaptive appearance $\mathbf{A_t}$ is a time-varying one that models all the shape-free textures $\mathbf{X}$ up to time $t-1$. We adopt the method described in [6] to build the shape-free texture $\mathbf{X_t}$, which is the 2D texture mapped on the ellipsoidal head model. At first, we project the ellipsoidal model onto the 2D image plane to get a 2D triangular mesh. Then the shape-free texture is created by texture warping from the triangular 2D mesh covering the face region in the input images using a piece-wise affine transformation $\mathbf{\Psi}$. Note that we utilize barycentric coordinate to shorten the computational time of the above warping process.

Mathematically, the affine warping function $\mathbf{\Psi}$ applied to an input image $\mathbf{Y_t}$ is denoted by:

$$\mathbf{X_t} = \mathbf{X}(\widehat{\mathbf{B}}_\mathbf{t}) = \mathbf{\Psi}(\mathbf{Y_t}, \mathbf{B_t}) \tag{1}$$

where $\mathbf{X_t}$ represents the shape-free texture patch and $\mathbf{B_t}$ is estimated head pose at time t.

To fully adopt statistical model, the graylevel of each pixel within $\mathbf{X_t}$ are assumed to be independent and can be modeled as Multivariate Gaussian Distribution: each pixel $\mathbf{x_i}$ has a corresponding mean value $\mu_\mathbf{i}$ and a deviation value $\sigma_\mathbf{i}$. This Gaussian assumption can partially explain the appearance variations caused by facial animation, illumination changing etc.. In summary, a multivariate Gaussian can represent the shape free texture vector $\mathbf{X_t}$:

$$P(\mathbf{X_t}|\mathbf{B_t}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{N}(\mathbf{x_i} : \mu_\mathbf{i}, \sigma_\mathbf{i})_\mathbf{t} \tag{2}$$

$\mathbf{N}(\mathbf{x_i} : \mu_\mathbf{i}, \sigma_\mathbf{i})_\mathbf{t}$ is the standard Gaussian distribution and n is the resolution of $\mathbf{X_t}$.

Given the Gaussian assumption, the cost function to stand for the discrepancy between $\mathbf{A_t}$ and $\mathbf{X_t}$ is selected as the Mahalanobis distance between them. Here, we introduce a confidence value to weight the contributions of mesh triangles. See [3] for its definition. Subsequently, the cost function is M-distance weighted a confidence factor:

$$e_m(\mathbf{B_t}) = \mathbf{min} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\frac{\mathbf{c_i}}{\mathbf{255}} \cdot \frac{\mathbf{x_i} - \mu_\mathbf{i}}{\sigma_\mathbf{i}})^\mathbf{2} \tag{3}$$

where $\mathbf{c_i}$ is the confidence factor weighting the triangle covering i^{th} pixel.

The above criterion is minimized using iterative first-order linear approximation. Expand $\mathbf{\Psi}(\mathbf{Y_t}, \mathbf{B_t})$ around $\mathbf{B_{t-1}}$ to a Taylor series and keep only its first-order gradient matrix $\mathbf{G_t}$. By approximating the current mean texture vector as $\mathbf{M_t} \approx \mathbf{\Psi}(\mathbf{Y_t}, \mathbf{B_t})$, the motion shift $\mathbf{\Delta B_t}$ can be expressed as:

$$\Delta\mathbf{B_t} = -\mathbf{G_t^\dagger}(\mathbf{\Psi}(\mathbf{Y_t}, \mathbf{B_{t-1}}) - \mathbf{M_t}) \tag{4}$$

where $\mathbf{G_t^\dagger}$ is the pseudo-inverse of $\mathbf{G_t}$. To gain more accuracy of $\mathbf{G_t}$, its numerical difference is listed. In practice, the minimization process is iterated till a 'local' minimal is arrived. See [6] for the registration proof and its gradient descent minimization process.

The outcomes of the above registration is the estimated head pose shift $\mathbf{\Delta B_t}$. We directly use it to update current head pose as:

$$\mathbf{B_t} = \mathbf{B_{t-1}} + \triangle\mathbf{B_t} \tag{5}$$

Once we get the current shape-free texture and register the estimated head pose, we can update the parameters in appearance At using the strategy of Online Appearance Models:

$$\mu_{i(t+1)} = (1-\alpha)\mu_{i(t)} + \alpha_{i(t)} \tag{6}$$

$$\sigma^2_{i(t+1)} = (1-\alpha)\sigma^2_{i(t)} + \alpha(x_{i(t)} - \mu_{i(t)})^2 \tag{7}$$

where $\alpha = 1 - exp(-\log(2/n_h))$ is a forgetting factor, which weights past observations in recursive filters, and nh represents the half-life of the envelope in frames (See [8] for details). All the model parameters in this recursive filter is learned online and the most recent information is retrieved frame by frame. In practice, equation(12) is not used until it reaches a stable variance. So for the initial n_i frames, a naive $\alpha = a/t$ is set to update variance (a makes sure $\alpha \in (0.01, 0.1)$).

To partially offset the influence of outliers caused by tracking failure or non-rigid face animation, we define a simple occlusion function. For the i^{th} pixel x_i, if $|\frac{x_t - \mu_t}{\sigma_t}| < \mathbf{c}$, then the pixel will be regarded as outliers; otherwise, it is inlier. For outliers, we don't update their means and variances. For inlier's pixels, we update them by equation (11), (12). In our experiments, $\mathbf{c}$ is set to be 3.

4.1.3 Drifting Correction and Re-initialization

As drifting problem accompanies all the online tracking framework, to define a strategy to correct it is indispensable. Iain Matthews [12] propose several strategies to partially correct drifting problems. We adopt a similar strategy. While the registration error $\mathbf{e_m}$ is bigger than a threshold value $\mathbf{e_{drift}}$, we say that drifting happens. Then, the template will be corrected by the initial template. As our template is defined as a multivariate Gaussian texture vector, the mean pixel value $\mu_{\mathbf{1:n}}$ in current appearance template $\mathbf{A_t}$ is the right one to be corrected:

$$\mathbf{A}^*(\mu_{\mathbf{1:n}})_\mathbf{t} = (\mathbf{1} - \theta \cdot \mathbf{A}(\mu_{\mathbf{1:n}})_\mathbf{t} + \theta \cdot \mathbf{A}(\mu_{\mathbf{1:n}})_\mathbf{0} \quad (8)$$

$\mathbf{A}(\mu_{\mathbf{1:n}})_\mathbf{0}$ is the initial template and θ is a weight value. Then the current image will be re-registered with the above corrected template. The re-estimated head pose is denoted as $\mathbf{B}^*_\mathbf{t}$, which is supposed to replace the previous $\mathbf{B_t}$.

Occasionally, due to global occlusions or too fast movements, the tracking will completely fail. Accordingly, the tracking must be re-initialized. We treat the registration error $\mathbf{e_m}$ as the norm to judge tracking failures. When it exceeds a certain value $\mathbf{e_{failure}}$, the tracking will be re-initialized by a face detecting process and re-registered strategy.

In Figure 3, we display some tracking examples respectively under challenging head pose variations, full occlusion, local appearance variation and varying illumination. The experimental results illustrate the effectiveness of our defined tracking framework.

4.2. View-Independent Video Texture

After determining head poses, the view-independent face textures could be extracted by warping current face regions. In practice, those textures are generated in the same way as the shape-free texture in head pose tracking in Section 3.3. The barycentric coordinates for each pixel is also pre-computed at the first frame. Subsequently, at next frames, due to the affine-transform-invariant property of the barycentric coordinate, the color value of each pixel is directly determined by using bilinear interpolations of the pixels in the mesh-covered face region. Figure 4 displays some sampled view-independent textures from two subjects.

Figure 4. View-independent facial textures extracted in two video clips.

5. Head Modeling and Video Mapping

Ahead of pose tracking and video mapping, we must build a personalized 3D head model. Even nowadays, a fully automatic head model adaptation is still an unsolved problem. In the paper, we develop an easy-to-use interface to support interactive head modeling. Figure 5 is a snapshot of a model adaptation result. The user interactively fits model keypoints to facial features given in the two photos. Then, the mesh surface is subdivided to get a rather smoothing, dense head model. Note that the two photos had better to be orthogonal, one frontal view, one profile view.

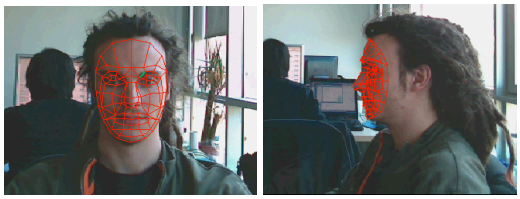


Figure 5. A snapshot of interactive head model adaption

Once head model is adapted and video textures are extracted, the online face avatar will be rendered by seamlessly mapping video textures onto the adapted head model. As head model is adapted, its texture coordinate is also modified to fit individual faces, which guarantee the correspondence between model keypoints and video textures. In practice, the avatar is rendered using OpenGL functions.

6. Experimental Results

The tracking framework is written in a non-optimized C/C++ program while video mapping is done using

OpenGL. The tracking speed is running around 15 FPS with setting K=8, shape-free-texture resolution to 40×42 on a PC with Xeon 3.0GHz CPU and 1.5G RAM. Actually, the computation of time-varying gradient matrix accounts for approximately 80% CPU time. The rendering speed is reduced to synchronize with the tracking speed.

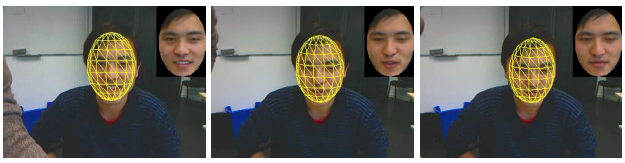


Figure 6. Simultaneous head pose tracking and view-independent texture extraction.

In Figure 6, we show a simultaneous head pose tracking and view-independent texture extraction example in a daily activity. Though there is only slight head motion, our program successfully tracks it and extracts accurate frontal-view textures. Note that in the textures, we lose gaze contact as eye pose is not in neutral view. Warping eye regions may correct this gaze contact problem.

Figure 7 demonstrates an example of a face avatar Avatar1. The video size is 640×480. The tracking initialization time is about 10s. The average tracking speed is 10fps while the max converge frequency is 30. Note that tracking speed is rather slow (around 4 fps) when drifting or occlusion happens. This is due to the extra time consumption of recovery and re-initialization process. The video mapping is almost seamless except in the nose part. It can be explained that the face textures are lack of depth information in nose regions. In general, the copied facial expressions is as realistic as the original video sequence. In particular, the real lip movements can be used for speed-synchronized animations.

In Figure 8, it is another avatar example Avatar2. The video sequence is captured in 640×480 and resized to 320×240 for tracking convenience. This is because that the computation of barycentric coordinates for view-independent textures at the first frame maintains more than ten seconds, which could reduce to about 5 seconds while half resized. However, we lose certain realism in a smaller image size, as shown in the figure. In this video sequence, lighting conditions are changing over time. As a result, light reflection in face textures is not uniform. In addition, the glasses and the forehead hair, detected as outliers, are not erased from the face textures. All these problems will be addressed in our future work.

A video demo containing more results can be found at http://www.youtube.com/watch?v=ZGwX00vtXH8.

7. Conclusion and Future Work

In this paper, we propose a new scheme to create online face avatar from monocular video sequences. In particular, the approach is targeted towards creating automatic, realistic and fully available avatar for game or collaborative work. Our approach is more efficient as it only needs tracking rigid head motion and directly uses texture mapping to reproduce facial expression. The encouraging results support the fact that a seamless mapping of video textures onto a static head model can produce the same realistic expression as dynamic facial animation does.

Its worthwhile to point out that there are still some limitations of the suggested scheme in practical online applications. Some improvements in our ongoing work are as follows:

Improve Tracking Performance The current tracking framework is not fast enough for real-time applications and is not robust enough to handle with very challenging video situations. One of our current work is to accelerate the tracking speed by combining the merits of some real-time algorithms [11] and GPU acceleration.

Texture Relighting and Deblurring The face textures extracted from video sequence are somewhat noisy. Lighting conditions are not constant through the whole videos and face colors are blurry due to the unexpected occlusions or motion blurring. Our future work is to relight and deblur extracted textures.

Gaze Enhancement For direct face mapping solutions, eye gaze awareness is often lost if face avatar is rendered to look at another target. One approach is to re-rotate head pose to remain eye gaze while the other is to track eye pose and warp eye regions to correct eye gaze. In the future, we will test both to enhance gaze for the online avatar.

References

[1] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross. Multi-scale capture of facial geometry and motion. *ACM Trans. Graph.*, 26(3):33, 2007.

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH'99*, pages 187–194, 1999.

[3] L. M. Brown. 3d head tracking using motion adaptive texture-mapping. *Computer Vision and Pattern Recognition*, 1:998, 2001.

[4] J. Chai, J. Xiao, and J. K. Hodgins. Vision-based control of 3d facial animation. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, July 2003.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE T-PAMI*, 23(6):681–685, 2001.

[6] F. Dornaika and F. Davoine. Online appearance based face and facial action tracking. *IEEE T-CSVT*, 16(9), 2006.

[7] J. Haber and D. Terzopoulos. Facial modeling and animation. In *ACM SIGGRAPH 2004 Course Notes*, 2004.

[8] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE T-PAMI*, 25(10):1296–1311, 2003.

[9] Y. Lee, D. Terzopoulos, and K. Walters. Realistic modeling for facial animation. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62, 1995.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[11] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135 – 164, November 2004.

[12] I. Matthews, T. Ishikawa, and S. Baker. The template update problem. *IEEE T-PAMI*, 26(6):810 – 815, June 2004.

[13] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesizing realistic facial expressions from photographs. In *SIGGRAPH'98*, pages 75–84. ACM, 1998.

[14] P. Quax, T. Jehaes, P. Jorissen, and W. Lamotte. A multi-user framework supporting video-based avatars. In *NetGames'03*, pages 137–147. ACM, 2003.

[15] B. Sumit, E. Irfan, and P. Alex. Motion regularization for model-based head tracking. *ICPR*, pages 611–616, 1996.

[16] C. T. V.Blanz and T.Vetter. Reanimating faces in images and video. *In EUROGRAPHICS*, 22, 2003.

[17] D. Terzopoulos and Y. Lee. Behavoiral animation of faces: Parallel, distributed, and real-time. *In SIGGRAPH 2004 Course Notes*, 2004.

[18] B.-J. Theobald, I. A. Matthews, J. F. Cohn, and S. M. Boker. Real-time expression cloning using appearance models. In *ICMI'07*, pages 134–139, 2007.

[19] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, 2005.

[20] J. Xiao, T. Kanade, and J. F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. In *International Conference on Automatic Face and Gesture Recognition*, page 163, 2002.

Figure 3. Head pose tracking examples under various conditions. From up to down, the respective conditions that each row displays are (1)challenging head movements;(2)face occlusion;(3)wearing on glasses;(4)varying illumination.

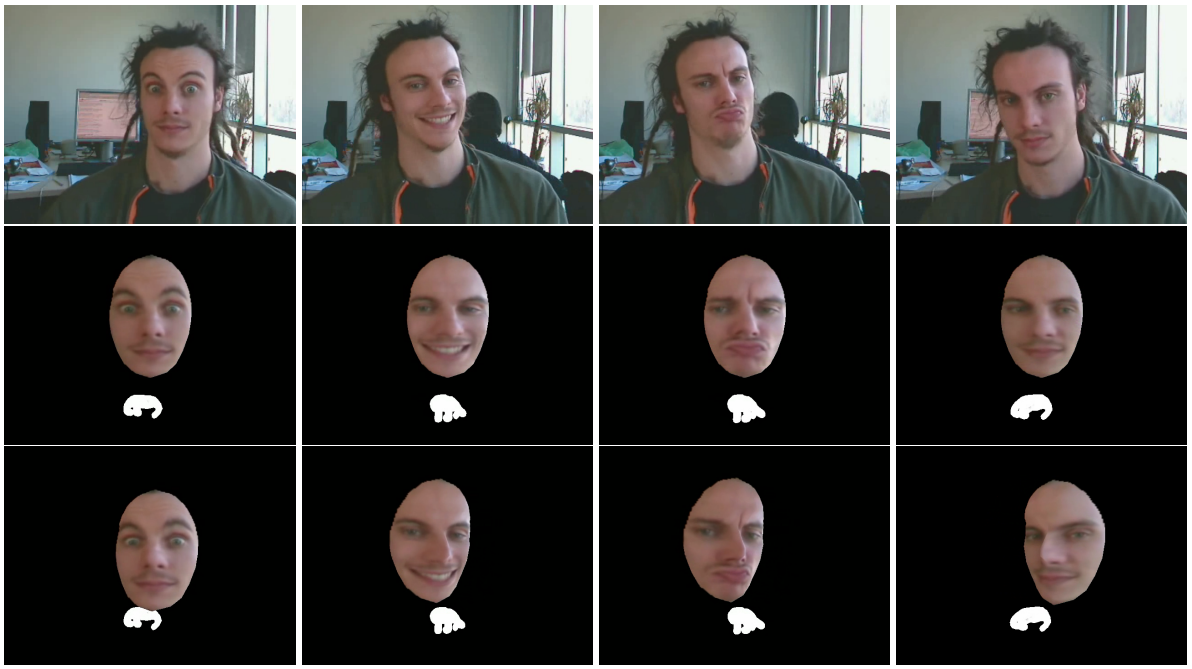


Figure 7. Face avatar Avatar2 with a white hand pointing out head pose. The first row images are chosen from a webcam-captured video; then in the middle row, his corresponding facial expressions are real-time rendered in a front view; The lowest row are the results of simultaneous facial expression and head pose reproducing.

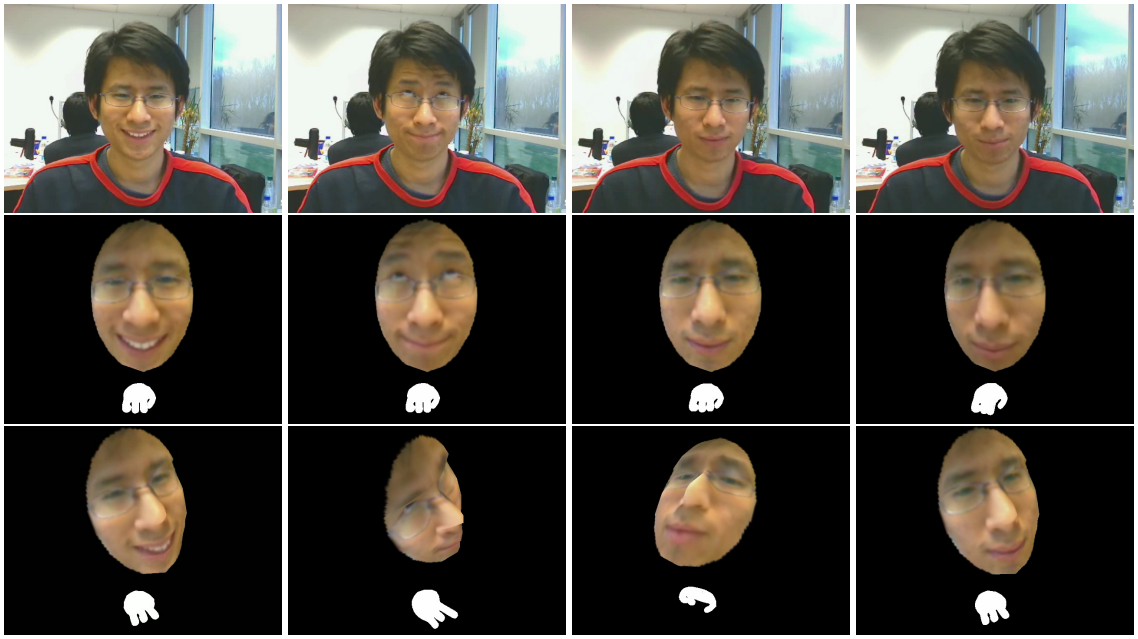


Figure 8. Face avatar Avatar2 with a white hand pointing out head pose. The upper row images is sampled from a video sequence; the middle row shows the reconstructed avatar under near-frontal head gesture; then in the lower row, the avatar is re-rendered into random gestures.

Individual Differences in Facial Expressions: Surprise and Anger in the Emotion Evoking Game

Ning Wang
Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Marina del Rey, CA USA
nwang@ict.usc.edu

Stacy Marsella
Information Science Institute
University of Southern California
4647 Admiralty Way, Marina del Rey, CA USA
marsella@isi.edu

Abstract

Studies of actual facial expressions can help improve the design of recognition algorithms by providing real data as well as uncover potential issues associated with using facial expression recognition to control a game. We have developed a computer game called the Emotion Evoking Game (EVG) that provides a systematic method for evoking emotion in response to events in game play.. In our work, we have been investigating the nature of the facial expressions evoked during game play in EVG. In this paper, we discuss a detailed analysis of facial expressions in response to situations intended to induce surprise and anger. Our data shows great individual differences in facial display of surprise and anger. These individual differences raise serious challenges to use of the recognition emotion and specifically the recognition of the expression of emotion on the face as means to control game play.

1. Introduction

With the advance of computer vision technology, researchers have begun to investigate means to improve human computer interaction using gesture and facial expression recognition. Facial expressions in particular are argued to be important indicators of one's emotional states. Ekman, Friesen, and Ellsworth [5][8] claim that facial expressions can provide accurate information about emotion, e.g. distinguishing between pleasant and unpleasant emotions. It's not surprising that video games, which create emotionally engaging interaction, have attracted great interest to use facial expressions to control gaming experiences [21].

Studies of actual facial expressions can help improve the design of recognition algorithms by providing real data as well as uncover potential issues associated with using facial expression recognition to control a game. For example, even though there are facial expressions of emotions that are universal [10], there are great individual differences in how these expressions are displayed. Cohn, Schmidt and Gross and Ekman [2] investigated individual differences in positive facial expression in 4 to 12 months and found individual differences in facial expression that are stable over time in both individual and interpersonal context.

The study of human emotions and emotional expressions can benefit from a systematic method for emotion evocation. Over the years, researchers have begun to experiment using computer games as a means to induce emotions. The use of computer games can induce task-related emotions such as frustration. It can also induce social emotions such as guilt, or anger due to betrayal of a teammate. One of the first games used to induce emotions is the Geneva Appraisal Manipulation Environment (GAME) – a Pac-Man like game running under DOS created by Geneva Emotion Research Group from the University of Geneva [13]. In GAME, events representing different values of appraisal dimensions are used to induce emotions in players. It demonstrated that computer games can be effective as an emotion inducing method for emotion and facial expression research. Kappas, A., Pecchinenda [14] used two games called aMaze and Playgame to manipulate appraisal of emotions. Similarly, van Reekum [18] used X-Quest to study appraisal dimensions. We have developed a computer game called the Emotion Evoking Game (EVG) [19] that provides a systematic method for evoking emotion in response to events in game play. EVG is designed to assist development and evaluation of new techniques for recognizing emotions and facial expressions. It allows researchers to systematically explore factors that elicit emotion. EVG can reliably induce emotions and facial expressions [19].

In our work, we have been investigating the nature of the facial expressions evoked during gameplay in EVG. An important aspect of studying facial expression is how to record them in detail. Some facial expressions are quite fleeting. Ekman [9] argues that micro-expressions that can be on the order of 40 ms. Facial expressions can also be subtle [9], with dynamic properties that can impact human interpretation [15]. To study human facial expression closely, we used a high speed camera to capture the richness and subtlety of facial expression at a

978-1-4244-2154-1/08/$25.00 ©2008 IE

fine grain level. Previous work [20] on the emotional reactions to a surprise situation in EVG revealed complex dynamics of the surprise facial expression over time. In this paper, we provide data on anger reactions along with additional result on surprise.

A key issue raised in this work concerns the pattern of facial display associated with these emotions. Darwin [3] suggested that surprise is a biologically determined facial display consisting of three components: eyebrow raise, widening of the eyes, and opening of the mouth/jaw drop. This corresponds to Inner & Outer Brow Raiser (AU1 & AU2), Upper Lid Raiser (AU5), Lips Apart (AU 25) and Jaw Drop (AU26) in the Facial Action Coding System (FACS) [6]. Other research argues that facial expressions of emotion are more often partial than complete [1][16]. Studies by Reisenzein [17] find that surprise doesn't correspond to the three component display model. In terms of anger, Ekman and Friesen [7] argue that anger facial expression is often associated with Brow Lowerer (AU 4), Upper Lid Raise (AU 5) and Upper Lip Raiser (AU 10).

Our work presented in this paper reveals considerable individual differences in the facial expressions in response to surprise and anger inducing situations. This raises serious challenges to the use the recognition of the expression of emotion on the face as means to control game play.

2. EVG: Emotion Evoking Game

EVG [19] is adapted from an open source game called Egoboo [12]. It is implemented as a role-playing dungeon adventure game. The current setup includes events targeted to evoke five different emotions: boredom, surprise, joy, disgust and anger, in order. The story in the current study is that the player, accompanied by a teammate (a non-player character), starts out in an underground palace with the goal to collect 2000 units of gold. In the end, the player defeats the enemies and successfully collects 2000 units of gold. Then the teammate betrays the player by killing him and stealing the gold. The five main emotion evoking phrases of this setup are discussed below.

Collection: First, the player and teammate go about the dungeon to collect gold by opening up chests placed in separate chambers. During this stage, there's no enemy presence. This stage is intended to evoke boredom.

Shock-and-Awe: The next stage starts when player walks into the last chamber and found himself under the attack from boss enemy and several other powerful enemies. Events at this stage would are designed to evoke surprise.

Victory: After battling with the enemies, player defeats all the enemies and the boss enemy drops 1000 units of gold, which can help the player achieve his goal. Events at this stage are hypothesized to induce joy.

Figure 1: Screenshot of Emotion Evoking Game (EVG).

Betrayal: While the player is collecting gold, teammate betrays the player by attacking him. Events at this stage are designed to evoke disgust.

Loss: Eventually the teammate kills the player and claims victory. Player loses all the gold collected. Events at this stage are targeted to induce anger.

3. Experiment

3.1. Participant

Thirty-five subjects (40% women, 60% men) participated in this study. They were recruited using an online local community website and were compensated $20 for one hour of their participation.

3.2. Procedure

Upon his arrival to the laboratory, the subject was told he was in a study to evaluate a computer game that's under development. Then the subject read and signed the consent form which includes agreement to be videotaped if they agree to participate.

Next the subject filled out the pre-questionnaire packet.

After finish filling out the packet, the subject was led into the computer room. The subject sat in the chair in front of the experiment computer and was encouraged to find a comfortable position so movements during the game would be minimal. At the same time, the experimenter adjusted the high speed camera. At this point, the computer display showed the welcome screen of EVG, which reads:

"Collect gold in the underground palace. Your goal is to collect 2000 gold. Your name is Louis. Alexis is your team member. Alexis can help you heal. Alexis has the key to the last chamber."

The experimenter went over the welcome screen with the subject make sure it's understood. Then the experimenter started the screen capture and the high speed camera. The subject started to play EVG and experimenter left the room.

After finished playing EVG, the subject was led to a side room. Experimenter presented with post-questionnaire packet and explained to the subject:

"Here are five copies of the same questionnaire. On each copy of the questionnaire, you may report an event or a moment while you were playing the game that you have felt emotions. Try to think of five events or moments that you felt emotions during the game."

Upon completion of the post-questionnaire packet, the experimenter thanked the subject for his/her time and participation, and paid subject $20.

3.3. Equipment

EVG was running on a DELL Precision 690 desktop with 19 inch display. Two speakers and a Saitek P2500 Rumble game controller were connected to this computer. Subject's interaction in EVG was captured using Fraps 2.8.2 at 60 fps. A Vision Research Phantom v10 camera was used to capture facial expression at 240 fps. It was connected to the experimenter's computer, which is a DELL Precision 650 desktop with 19 inch display. To produce enough light for the camera, the computer room was lit by 15 floor lamps with 3 florescent light bulbs on each lamp. Each bulb is equivalent to 100 Watt. All the lamps were facing the wall to diffuse the light.

3.4. Measure

Only two minutes of subject's facial expression (last two minutes before the game ends) was captured due to memory limits of the high speed camera. After the experiment, the research team synchronized the game play video and the facial expression video. A certified FACS coder from the research team viewed the synchronized video and marked appearance of different Action Units right after Shock-n-Awe and Loss event happens.

4. Results

4.1. Analysis of response to Shock-n-Awe event

Data from 6 subjects are excluded due to technical difficulties. As a result, data from 29 subjects are reported.

To conduct detailed analysis of the facial expression, the FACS coder coded the appearance and timing of the action units in the facial expressions. In the post-questionnaire, 65.5% of the subjects reported feeling surprise at Shock-n-awe, 20.7% showed eyebrow raise, 41.4% showed mouth open or jaw drop, and 17.2% showed widened eyes. In subjects who reported feeling surprise (n=19), 47.4% showed surprise facial expression, 21.1% showed eyebrow raise, 52.6% showed mouth open or jaw drop, 10.5% showed widened eyes, no subject showed all of the three components and 47.4% showed none of the three components.

Figure 2: Subjects' facial expression right before and after Shock-n-Awe event.

We observed great individual differences in the display of surprise facial expression. As noted earlier, our data showed that 47.4% of the subjects who reported surprise didn't show any of the three components proposed by [3]. Even when any of the three components is displayed, the facial expression often consists of only one or two, instead of all of the three components. Figure 2 shows some additional individual differences in the surprise facial expression. Some showed opened mouth and/or jaw drop with raised eyebrows while some showed opened mouth and/or jaw drop with tightened eyebrows (5th row, subject on the left in Figure 2).

Some showed tightening of the eyes. Some showed wrinkling of nose (AU9), which is an indication of disgust. Interestingly, some reacted with puckered lips (4th row, subject on the right in Figure 2) or tightening of the neck (2nd row, subject on the left in Figure 2), while others showed very little facial expression change. The variability of the facial expression reacted to Shock-n-Awe could be that instead of feeling surprise alone, subjects felt a blend of emotions, such as fear or disgust in addition to surprise. In addition to the immediate response, we also found great variance in the dynamics of facial expression following surprise. Some subjects followed the surprise with a smile. Some subjects ended the surprise facial expression with pulling only one side of the lip corner upwards. Some followed the surprise with an expression appears to be anger then fear.

Figure 2 also shows the display of action units that are normally not associated with surprise expression. For example, the subject in the first row on the right showed chin raise (AU 17) and funneled lips (AU 22). We observed similar lip funneling movement from 4 other subjects as well. This behavior was shown repeatedly when subjects were fighting with their enemies. It was as if the subjects were trying to control their character using their lips. Interestingly, we only observed this behavior in male subjects.

We also see a lot of frowning (AU 4) at this stage. However, most of the subjects showed frowning before the Shock-n-Awe event. This may be an indication of engagement. It could also be caused by the bright light in the room for the high speed camera.

4.2. Analysis of overall response to Loss event

To study the facial expression in response to the Loss event, video sequences of reactions to Loss event from 17 subjects were analyzed. From the self-report, 29% of subjects reported feeling of surprise and only 24% of subjects reported feeling of anger. Later, two human raters viewed the videos and rated the emotions revealed by the facial expression ($r = .483$). In the 17 video sequences, 52% of them showed anger according to the two human raters. The FACS coder then coded the appearance and timing of the Action Units in the facial expressions. Table 1 shows the frequency of some of the Action Units we observed. Action Units associated with head and eye movements, such as head up, head down, eyes left and etc, were observed but not listed in this table because they are not the focus of this paper. The third column of Table 1 shows the frequency of Action Units that occurred across all facial expressions in reaction to Loss event. The fourth column shows the frequency of Action Units that occurred in a facial expression that's judged as anger by the two raters. From the Table 1, we can see that the most frequently displayed Action Units are Lips Apart (AU 25) and Jaw Drop (AU 26). However, we need to take into account that over half of the time, both AU 25 and 26 were already present before the Loss stage. In those cases, the appearance of AU 25 and 26 are triggered by events prior to the Loss event, such as events in the Shock-n-Awe or Betrayal stage. In addition, the presence of AU 25 and AU 26 could mean that subjects were feeling surprise as well. In the video sequences in reaction to the Loss event, we found a strong correlation between surprise and AU 25. All except one of the facial expressions coded with surprise showed both AU 25 and 26. The only one surprise expression that's not coded with AU 25 and AU 26 is coded with AU 72, which means lower face not visible.

Table 1. Frequencies of Acton Units in the facial expressions in response to the Loss event.

Action Unit		Frequency (%)	
Number	Name	Overall (Total: 17)	Anger (Total: 9)
1	Inner Brow Raise	29	22
2	Outer Brow Raise	29	22
4	Brow Lower	47	67
5	Upper Lid Raise	12	11
6	Cheek Raise	41	44
7	Lids Tight	6	0
9	Nose Wrinkle	24	11
10	Upper Lip Raiser	29	44
12	Lip Corner Puller	47	56
25	Lips Part	76	78
26	Jaw Drop	82	89

From Table 1, we can also see that some Action Units are more strongly tied to anger compared to the others. For example, the frequency of AU 4, the Brow Lowerer, increased 20% in an angry expression. This shows that AU 4 and 10 can be indications of anger. This is consistent with the categorization of Action Units suggested in the FACS manual [7].

Nine of the facial expressions shown by subjects were rated as anger. From Table 1, we can see that the most frequently occurred Action Units in anger facial expressions are AU 4, 6, 10 and 12. Overall, 89% of the anger expressions in our data consist of either AU 4 or AU 10. And 44% of the anger expressions include AU 4 and AU 12. A combination of AU 4, 6 and 12 were shown 33% of the time. Angry expressions from four subjects were shown in Figure 3. The first and the last subjects showed both frowning and smiling. We interpret them as masked anger. The subject in upper right showed a rather "contained" anger, while the subject in lower left showed a more pronounced anger. However, showing tongue is a rather rare anger

expression in the western culture. We only observed one instance in our data.

Figure 3: Subjects' facial expression before and after the Loss event.

In general, our data showed two kinds of anger facial expression. The first one is a combination of AU 4 and AU 10. This is more in line with the typical anger expression. Another variation of the anger expression is the combination of AU 4 and AU 12. This is an interesting departure from the typical anger facial expression. AU 12, which is often associated with smile, is not considered as part of the characteristics of anger expression [7]. In our data, we found that AU 4 is generally rather brief when combined with AU 12. And it occurs either before or while AU 12 is in action. This combination of Action Units indicates a display of frowning briefly before or during a smile. When frowning happened before a smile, it could mean that the subject felt angry and tried to cover it up with a smile. When frowning happened during a smile, it may reveal a "leakage" of anger from the smile displayed on the surface.

5. Discussion

In this paper, we presented a detailed analysis of facial expressions induced by the events in the Shock-n-Awe and Loss stage in the current implementation of EVG. Overall, our data seems to support the idea that surprise facial expression is often partial rather than complete [17]. None of the subjects showed all three components of surprise facial expression specified by Darwin [3]. Some action units that are not typically associated with surprise were also observed. This indicates considerable differences across subjects in terms of the emotions evoked and expressed. In terms of the reaction to anger event, 52% of the subjects displayed anger according to the human raters. However, anger facial expression was displayed very differently across all the subjects. Majority of the anger expression includes lowering the eyebrow (AU 4). Some subjects showed wrinkled nose (AU 9) which is often considered as an indication of disgust. Prior to the Loss stage, the friend character betrayed the subject and began to attack the subject. This event is designed to induce disgust towards the friend. So, it was possible that the feeling of disgust was still lingering during the Loss stage and resulted in display of wrinkled nose. In addition, the events in the Loss stage, e.g. killed by the friend, were direct consequences of the friend's betrayal. It's quite likely that subjects felt a blend of anger and disgust towards the so-called friend.

Interestingly, even though anger was displayed by over half of the subjects, AU 12, the lip corner puller, was also shown about 50% of the time. AU 12 is often associated with smile and considered as an expression of joy. It is an Action Unit that can be controlled voluntarily. The co-presence of smile and negative emotions such as anger and disgust indicates that subjects were possibly following the display rules to cover up socially undesirable emotions with smile. An experiment done by Ekman and Friesen [4] found that the Japanese students would mask expressions of negative emotion with a polite smile when a person in authority would be present much more than American students. In our study, the experimenter can be considered as an authority. Subjects were perhaps trying to use smile to cover up the anger and disgust they felt. Alternatively, subjects may simply find it funny that the friend turned on their character in EVG. So, they may felt happy and disgust simultaneously.

One of the limitations of the study is that only one FACS coder coded the facial expression in the video. Using at least two coders could increase the coding reliability.

Overall, we found EVG was very successful in evoking emotions and a wide range of facial expression. Research on facial expression recognition could benefit from tools like EVG. In addition, we found considerable variability in facial display of surprise and anger. These individual differences created great challenge in the design of games that uses facial expression to control the game experience. The study presented here showed the analysis of surprise and anger facial expression across different subjects. Perhaps further analysis of the facial expressions for each individual subjects over a period of time can help shed light on the more stable individual differences. These longitude studies can help fine tune the facial expression recognition algorithms to adapt to these individual differences.

References

[1] Carroll, J. M., Russell, J. A.: Facial expressions in Hollywood's portrayal of emotion. Journal of Personality and Social Psychology, 72, 164-176. (1997)

[2] Cohn, J.F., Schmidt, K., Gross, R., Ekman, P: Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform

person identification. Fourth IEEE International Conference on Multimodal Interfaces, 491-496. (2002).
[3] Darwin, C. The expression of the emotions in man and animals. London: Murray. (1872)
[4] Ekman, P., Friesen, W.V.Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17, 124-129. (1971)
[5] Ekman, P., Friesen, W. V., Ellsworth, P. Emotion in the human face. Elmsf Pergamon. (1972)
[6] Ekman, P., Friesen, W. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
[7] Ekman, P., Friesen, W. V.: Investigator's guide to the Facial Action Coding System. Palo Alto, CA: Consulting Psychologist Press. (1978)
[8] Ekman, P.: Emotion in the human face. New York: Cambridge University Press. (1982)
[9] Ekman, Paul. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. New York: Norton. (1985)
[10] Ekman, P.: The argument and evidence about universals in facial expressions of emotion. In H. Wagner & A. Manstead (Eds.), Handbook of social psychophysiology (pp. 143-164). Chichester: Wiley. (1989)
[11] Ekman, P., Rosenberg, E. What the Face Reveals. New York: Oxford University Press. (1997)
[12] Egoboo. http://zippy-egoboo.sourceforge.net/ (2000)
[13] Kaiser, S., Wehrle, T.: Situated emotional problem solving in interactive computer games. In Frijda, N.H., (ed.), Proceedings of the VIXth Conference of the International Society for Research on Emotions, 276--280. ISRE Publications (1996)
[14] Kappas, A., Pecchinenda, A. Don't wait for the monsters to get you: A video game task to manipulate appraisals in real time. Cognition and Emotion, 13, 119-124. (1999)
[15] Parkinson, B., Fischer, A. H., & Manstead, A. S. R. Emotion in social relations: Cultural, group, and interpersonal processes. New York: Psychology Press. (2005)
[16] Reisenzein, R.: Exploring the strength of association between the components of emotion syndromes: The case of surprise. Cognition and Emotion, 14, 1-38. (2000)
[17] Reisenzein, R., Bördgen, S., Holtbernd, T., Matz, D.: Evidence for strong dissociation between emotion and facial displays: The case of surprise. Journal of Personality and Social Psychology, 91, 295-315. (2006)
[18] van Reekum, C. M. Levels of processing in appraisal: Evidence from computer game generated emotions. Doctoral dissertation, University of Geneva, Switzerland. (2000)
[19] Wang, N., Marsella, S. Introducing EVG: An Emotion Evoking Game, Intelligent Virtual Agent (2006)
[20] Wang, N., Marsella, S. Individual Differences in Expressive Response: A Challenge for ECA Design. Autonomous Agents and Multiagent Systems (2008)
[21] Zhan, C., Li, W., Ogunbona, P., Safaei, F.: A Real-Time Facial Expression Recognition System for Online Games. International Journal of Computer Games Technology, vol. 2008, Article ID 542918, 7 pages, (2008)

A Mimetic Strategy to Engage Voluntary Physical Activity In Interactive Entertainment

Andreas Wiratanaya
Gouleystrasse 122b
52146 Würselen
Germany
wiratanaya@gmail.com

Michael J. Lyons
College of Image Arts and Sciences
Ritsumeikan University
56-1 Tojiin Kitamachi Kita-ku
Kyoto Japan
lyons@im.ritsumei.ac.jp

Abstract

We describe the design and implementation of a vision based interactive entertainment system that makes use of both involuntary and voluntary control paradigms. Unintentional input to the system from a potential viewer is used to drive attention-getting output and encourage the transition to voluntary interactive behaviour. The iMime system consists of a character animation engine based on the interaction metaphor of a mime performer that simulates non-verbal communication strategies, without spoken dialogue, to capture and hold the attention of a viewer. The system was developed in the context of a project studying care of dementia sufferers. Care for a dementia sufferer can place unreasonable demands on the time and attentional resources of their caregivers or family members. Our study contributes to the eventual development of a system aimed at providing relief to dementia caregivers, while at the same time serving as a source of pleasant interactive entertainment for viewers. The work reported here is also aimed at a more general study of the design of interactive entertainment systems involving a mixture of voluntary and involuntary control.

1. Introduction

Human-computer interfaces employing face and body gestures as input may be classed into two broad categories: those which take intentional, conciously expressive input from the user and those which process spontaneous, natural actions made without deliberate intention. In the past decade or two, there has been a tremendous interest in the latter class of human-machine interfaces. In particular, the automatic recognition of facial expressions has become a major sub-topic of computer vision and pattern recognition research, as has been discussed in depth in the widely cited reviews of the field [1,2].

The former class of interfaces, designed for voluntary control, have attracted increasing attention more recently, stimulated by the widespread availability of hardware capable of handling real time face and body gesture recognition, as well as efforts to develop methods of human computer interaction which go beyond the keyboard and mouse. To give a specific example of this kind of approach, a series of such works involving intentional action of the face and mouth in voluntary real-time interactive control is summarized by Lyons [3].

While the distinction between voluntary and involuntary interaction may be useful one for system designers, in fact there is no strict dichotomy between these two forms of gestural interaction, but rather a maleable continuum that depends on a user's level of awareness and expertise. Consider, for example, interaces aimed at the processing and interpretation of expressive movement in dance and musical performance, such as is treated in the work of Camurri *et al.* [4], using the EyeWeb platform. It is not difficult to conceive of situations in which a performer or visitor to an installation may be initially unaware that their movements are being tracked and used to control sound synthesis, lighting, or other effects. Perhaps more commonly, a performer may be initially unfamiliar with the properties of a given interactive system and not capable of fully intentional control. With increasing experience and understanding of a system, a user may develop control intimacy [5], whence the mode of interaction will shift from having the character of involuntary recognition of actions, to an expert voluntary 'playing' of that system, just as a skillful musician plays a musical instrument.

In the work reported here, we considered a further possible blending of voluntary and involuntary modes of interaction to design and implement a system which encourages a user to make the transition from a passive, involuntary state, to a physically active, voluntary forms of interaction. In other words, we are interested in developing systems which can capture a subject's attention, engage their interest, then encourage them to play in a voluntary and active fashion. The specific prototype we describe was developed within the context of a larger project aimed at assisting the caregivers and family members of late-stage dementia patients [6, 7]. The

Figure 1: With the iMime system non-verbal interaction is driven by data about a subject's gaze, gestures, and motion qualities as observed using video input from multiple cameras.

general theme of this project is to develop entertainment interfaces suitable for dementia sufferers which will enjoyably hold their interest, thus providing occasional relief for care providers who need periods of time away from fully focussed care in order to attend to other pressing matters.

While various prostheses exist to help people with physical impairments, the development of technology for impaired cognitive abilities has only recently become the topic of active research [6]. In the case of dementia care the requirement for constant attention can create an unmanageable burden for the patient's family members [7]. The resulting stress can, in turn, have negative effect on the patient's well being. One strategy for reducing the burden of care, as well as the resultant stress, is to keep the patient entertained with audio-visual media that capture and hold the attention for even a limited period of time.

Recently Kuwabara et al. [7] presented a general framework for providing online support for people with dementia or severe memory-impairment giving the concrete application scenario of reminiscence videos which present elderly people with personalized memory stimulating images from their past. To add user interactivity to this application, Utsumi *et al.* [8] explored a content switching strategy for attracting and maintaining the attention of video watchers. This uses the subject's gaze direction as a measure of attention, switching to a different channel whenever the patient starts to lose interest. We have substantially extended this approach by designing and implementing a novel interactive interface with response-dependent adaptation of content display. Instead of reminiscence video contents we are exploring a much more interactive paradigm which uses a real-time lifelike animated character. Within this domain there has been extensive study of embodied conversational agents [9], however there has comparitively little study of purely non-verbal animated characters. Guided by a clinician familiar with the special needs of late-stage dementia sufferers, we have designed a framework for the interaction between a human and a virtual character. Since middle to late stage dementia patients often suffer from severely impaired capacity for verbal communication we base our system primarily on visually-drive non-verbal interaction. This has led us to consider the metaphor of the non-verbal performance of a mime. As can be commonly observed in street performances, mimes are well versed in non-verbal communication. We developed a system, called *iMime*, which draws inspiration from this ability of skilled mime performers to attract and hold attention, and entertain, using non-verbal interactive behaviour. Further support for our approach may be found in the study of Bailenson and Yee [10] in which embodied agents mimicking a viewer's head movements were found to be more persuasive than those which used recorded movements. While the prototype we describe here is still some stages removed from what can be applied in a clinical setting, or the home, valuable lessons have been learned from the design, implementation, and preliminary testing of iMime system.

2. System Design

Figure 2 shows a schematic of our iMime system. It illustrates the interaction flow between a human subject and a virtual character generated by a real-time animation engine. Our design is motivated by the performances of street mimes. In street mime a performer often makes a few stylized or humorous actions then freezes. To prompt further motion an observer is expected to engage in some form of active response: either by making a financial donation or reacting to the mime's behaviour. With this simple interaction strategy, a mime is able to bootstrap a completely non-verbal dialogue with strangers without recourse to speech.

We implemented this interaction metaphor as follows: a subject's movement is recorded by multiple video cameras at different scales. Computer vision algorithms are used to analyze the appearance and movement of the subject and draw conclusions about his/her attentional state. This information is passed to a state machine, updated with online reinforcement learning, to determine which behavior the virtual mime should exhibit next. In our prototype a set of animations were designed to be "mimesque", that is, to be entertaining and to encourage the viewer to respond with a reaction that in turn serves as an input to the vision system, hence closing an interaction loop.

2.1. Sensory Input

While a variety of sensors can be used to capture information about the attentional state of user, it was decided to restrict input to non-verbal and non-intrusive communication channels. In our prototype we use two video cameras to capture views of the patient at different scales: one camera is focused on the face while the other one captures a view of the entire upper body.

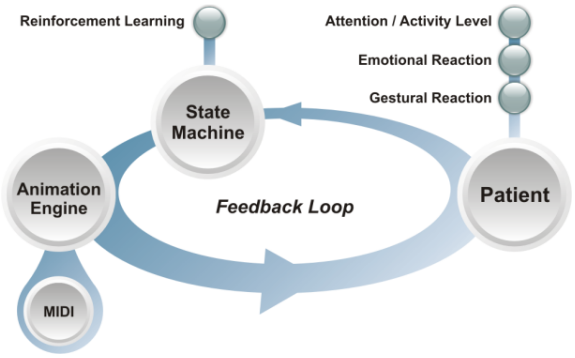


Figure 2: iMime system schematic.

2.2. Evaluating Attention

Humans show different signs of attention depending on their level of interest, ranging from simple observation over mild interest up to rapt attention. A subject passively viewing at the screen will merely look at the presented animation, whereas a deeply interested user may show unconscious facial reactions. An actively engaged user may further respond with voluntary body gestures. The vision system implemented for the iMime prototype is capable of: (a) determining whether or not the user is looking at the display, (b) recognizing the subject's current head orientation (c) classifying the subject's overall body motion (d) recognizing emotion primitives such as smile or frown (e) recognizing various basic gestures.

2.3. Adapting Character Behaviour to the Viewer

A general limitation of scripted animation systems or video contents is repetitiveness. Even the best scripts become boring after some time if the subject observes a non-changing, repeating pattern. Street mimes observe the reaction of the audience and adapt their behavior appropriately based on experience. It is interesting to notice that a mime will often show an act which by itself is not perceived as being funny or interesting but can be very entertaining in combination with later acts. We simulate this decision process by equipping the state machine that controls the animated character with an online reinforcement learning system. This system analyzes the attentional reaction of the viewer and devises a strategy to
to maximize that viewer's attention. Instead of greedily choosing the locally best solution the system is able to select behavioural states which could lead to a better solution in the future even if this involves taking a locally sub-optimal path.

2.4. Additional Input Channels

A mime uses different means to entertain her/his audience. While the mime is usually mute during the performance sometimes music is used to augment the gestures and expressions. We included an interface in our iMime prototype that allows for controlling the facial expression of the animated character by playing music files in MIDI format.

3. Implementation

3.1. Animation Engine

A mime conveys information using the two primary non-verbal channels: facial expression and body language. This has to be kept in mind when choosing a model for the animated character. Through extensive consultation with a clinician involved in research on dementia care, we learned that overly realistic human models would most likely not be acceptable. Animal models, while cute, impose limits on the usable range of body language. Finally, we settled on a relatively abstract model, shown in Figure 3, for its sufficient degree of expressivity, flexibility, and cartoon-like character.

3.2. Facial Animation

Facial animation is a widely studied topic in the field of computer graphics and several prior approaches are available. We first implemented a parameterized muscle model. The results were not sufficiently expressive and, in addition, controlling muscle parameters proved to be unintuitive. We therefore decided to use an alternate technique based on handcrafted morph targets. This method decomposes facial expressions into different primitives such as eyebrow raises, mouth shapes, or tongue positions. Figure 4 shows some examples. Blending expression primitives allows for the synthesis of a large variety of composite expressions. The curent model uses 42 different primitives and core expressions.

3.3. Body Animation

Limb movements typically involve non-linear rotations. So, the linear morph approach used for facial expression synthesis is not appropriate for synthesizing body poses.

Figure 5: Facial cues used by iMime: Left: face position detection. Center: classifying non-rigid motion using optical flow. Right: attention classification using the symmetric difference image.

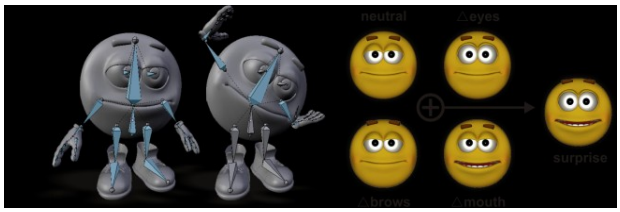


Figure 4: Structural components of the animation engine. Left: the skeletal system used for animating the body. Right: several facial expression morph targets.

Still, we would like to be able to combine movement primitives to generate variety. A widely used solution to this problem is the skeletal animation approach. A skeleton consisting of a hierachical arrangement of bones, inspired by the human body, augments the model. Moving the upper arm then move the lower arm, which in turn moves the hand. Each bone is assigned a region of influence in the model. Figure 4 shows the skeleton used to animated our character. With this approach, Animations can be parameterized and blended easily given the rotation angles for each bone. Adding Perlin noise [11] to each animation parameter, via convolution, gives a more lifelike appearance to the character.

3.4. Vision System

Our prototype uses two cameras to capture the patient at different spatial scales. One camera, aimed at the face, captures emotional and attentional information, while the camera focussed on the upper body extracts information about upper body gestures.

3.5. Analyzing the Face

We extended a system previously developed by our group [12,13] to analyze information visible in the viewer's face. The system combines automatic face detection and optical flow to classify facial expressions and to discriminate between rigid and non-rigid movement of the head.

For each frame we start by detecting the position of the face using the method of Viola and Jones [14]. We divide the face rectangle into seven regions of interest as shown in Figure 5: eyes, eyebrows, cheeks and mouth. In each region we compute the optical flow over previous frames [12]. Facial movement patterns can then recognized by classifying the characteristic flow vector comprised of the average flow vector for every region. We use patterns of optical flow to discern between non-rigid motion caused by facial expressions and rigid motion caused by head movement [13]. Translation of the head is detected by examining the motion of the face rectangle over the preceding frames. More generally, an analysis of the distribution of the main peaks in the flow field yields a good rigid motion criterion: a wide spatial distribution of motion peaks over the entire face indicates rigid motion. Similarly high average flow in all seven zones indicates presence of overall motion of the face.

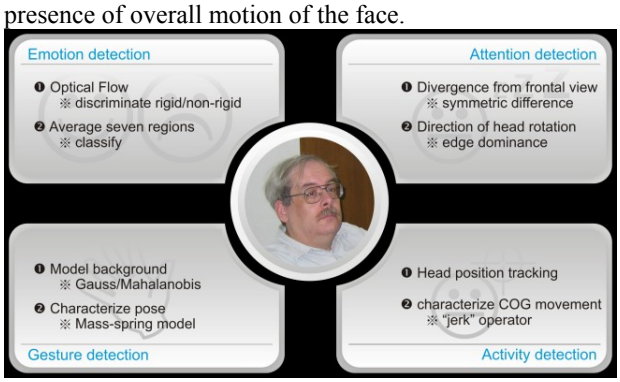


Figure 6: Overview of the vision system

To determine whether the patient is looking at the camera we exploit the symmetry of the human face. The face rectangle returned by the face finder is centered very precisely on the face. A good estimate of the face orientation can be obtained by computing the pixel distance between the left and right halves of the face after applying a median filter to the image. This is illustrated in Figure 5. We also employ an alternative method to determine left/right head orientation. If the head is turned to the left, the face detection rectangle moves so that one side of the rectangle includes the profile of the face, while the other side of the rectangle contains a relatively smooth area. In this case, there is a stronger presence of visual edges in the left side of the rectangle. Computing the center of gravity of all significant edges in the face region therefore provides an indicator for head orientation. Combining these two techniques described above yield a robust estimator of head orientation: while the first algorithm assumes roughly uniform lighting of the face, the second algorithm is insensitive to moderate lighting variations.

Finally to get an estimate of the spatial movement of the viewer we compute the fourth derivative ("jerk" operator) of the head position.

3.6. Analyzing the Upper Body

To classify the patient's pose we first separate the foreground pixels (patient) from the background pixels. The envisioned application of the system is in an indoor environment and we use a Gaussian model [15] to

describe the background, in which each pixel is described statistically. Figure 7 shows the typical results after removing the background from the input image based on the Mahalanobis distance from a threshold. The algorithm then binarizes the image the current and a mass-spring model [16] is used to recover the characteristic shape of the current pose. This process can be compared to dropping a piece of cloth from the top of the image onto the foreground object. Gravity pulls the cloth down over the object, which holds it in place. A converged drape of the body outline is shown in the right image of Figure 7. The algorithm returns a vector of height values that are normalized and correlated to previously acquired reference poses for classification.

3.7. Learning Module

A further novel feature of the platform we studied is that it attempts to learn from interactions with viewers in order to increase viewer attention. Online, real-time reinforcement learning on the attentional behaviour of the viewer is used to update the transitions between states of the animated character in such a way as to increase it's attractiveness.

We implemented a reinforcement learning system which consisting of two components: a *model estimator* to compute transition probabilities between Markov states based on analysis of human behavior data provided by the vision system, and a *value estimator* which computes the optimal policy based on the current state of the model estimator. The reinforcement-learning problem is equivalent to making optimal decisions in a Markov decision process, in which the world can be summarized in terms of its current state. In each state an agent is allowed to choose an action. Based on the action choice, the agent will enter a new state.

In essence, the learning algorithm rewards the adaptive system for capturing the interest of the subject. A simple operational definition of the viewer's "interest" is the time spent looking at the animated character. More concretely, looking at the animated character yields a reward of 1, whereas looking in a different direction gives a reward of 0. In practice, the reward is calculated at the same time as the adaptive agent is trying to decide which action to take next. This gives the viewer time to react to a change in state of the animated character's behavior, and allows the adaptive system to properly calculate changes based of the viewer's response. The job of the adaptive agent is to increase the reward, and thereby the user's attention.

3.8. Model Estimator

Suppose that the interactive character is a given state, for example the action "Routine 1" with the viewing attending, and chooses a certain action, for example

Figure 7: Gesture analysis. Left: input image. Center: background removal. Right: draping the outline.

switching to "Routine 3". Two outcomes are possible, the viewer will either continue to pay attention to the character or not. We can statistically model the observable outcome as a binomially distributed random variable, and estimate the distribution parameters using Bayes law with uniform priors. The *Model Estimator* of the reinforcement-learning module does exactly these tasks. For each possible state of the human-machine system *(Routine number, Viewer Attending or not Attending),* we need to estimate the parameters that describe this binomial distribution. The model estimator consists of a table of outcomes, representing the statistics for each binomial distribution – namely, the number of times the subject stopped paying attention after a particular change from one animated routine to another, as well as number of times the subject started paying attention to the animated character after such a change. This allows us to estimate the Markov transition probabilities as the maximum a-posteriori parameter values for each binomial distribution.

3.9. Value Estimator

The *Model Estimator*, at any given point in time, estimates the transition probabilities for the Markov decision process used by the adaptive animated character when switching behavior routines. Recall that the *Value Estimator* estimates a value function for a given set of model parameters. This value function, once determined, allows us to select the optimal action for the animated character. The popular Q-learning algorithm [17], running in a separate thread, allowed efficient real-time update of the value function. Whenever the model estimator is updated, based on the most recent human-machine interaction, the value estimator is updated to reflect the new model. Thus our adaptive character animation engine always makes optimal decisions based on experience interacting with the viewer.

3.10. Action Policy

We programmed iMime to change its action, with a 50% probability, every two seconds. Previous transitions are recorded, as is the viewers state of attention, or inattention. We use this data to update the model estimator module of the learning algorithm. With the model update and value estimator update, iMime looks at the optimal choice and selects it with a probability of $1 - \varepsilon$. It may also select any of the other, non-optimal, choices, with a

probability of ε. This procedure is known as the "epsilon-greedy" policy [17]. In this work, a value of ε = 0.125 was used.

4. Preliminary Results and Discussion

In order to get a first impression of the possible interactions, we have integrated all components into a prototype system with several video-based interaction modes. The iMime system is aware of user presence. If no user is seen the iMime animated character will either let his gaze wander around the room, showing randomly generated idle body movement or sit down bored and impatiently drum on the floor with his fingers. When a person enters his field of vision he will track him with his eyes and beckon to him. While the user is approaching, the character will indicate by gestures exactly how far he should stand from the camera in order to give a good view of the face. During interaction the character mimics facial expressions such as a smile or eyebrow movement and gestures such as waving. If the user does not interact for a certain period of time the character will first visibly ponder, then point at him and display an example gesture followed by a reward animation if it is mimicked by the viewer. iMime displays a scolding animation if the gesture is not mimicked by the viewer. If the viewers movements are erratic and rapid, iMime stops what he is doing, and gives a puzzled look while scratching his head. These behaviors are intended to be generic and intuitively understandable without verbal explanation.

The state transition machine is designed specifically to enable the transition of passive observation by the subject to active participation in the form of physical gesturing and movement of the subject as s/he attempts to elicit a novel behaviour from the virtual mime. Specifically the system instantiates a simplified non-verbal version of the children's game 'Simon Says': mimicking the gestures of the animated character result in a reward response while the subject is scolded if the response is incorrect.

We have observed the transition from passive to active interaction in informal preliminary tests of the prototype with naïve subjects, however further refinement of the system is necessary before further more definitive tests can be conducted.

5. Outlook

This paper described the design and implementation of a character animation system intended to attract, entertain, and engage viewers in a purely non-verbal way. Currently, all core modules are operational and have been integrated into a functional prototype. Preliminary tests with naïve users showed that the system is stable and works as intended. The reinforcement-learning module, however, adapts rather slowly in response to user behaviour. Future work involves investigating strategies for improving the rate of adaptation and, once this is accomplished, conducting further tests with subjects. While the system was designed in the context of a project aimed at dementia care, our more general intention is the development of a system which can serve as an effective testbed for studies of non-verbal interaction in encouraging active volutary physical involvement of users.

6. Acknowledgements

We thank Nick Butko for his contributions to the learning module, and Kiyoshi Yasuda for helpful discussions.

References

[1] M. Pantic and L. J. M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12): 1424 – 1445, 2000.

[2] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition* 36(1): 259-275, 2003.

[3] M. J. Lyons. Facial Gesture Interfaces for Expression and Communication. *IEEE International Conference on Systems, Man, and Cybernetics* Vol. 1: 598-603, 2004.

[4] A. Camurri, S. Hashimoto, M. Ricchetti, A. Ricci, K. Suzuki, R. Trocca, and G. Volpe. EyesWeb: Towards Gesture and Affect Recognition in Interactive Dance and Music Systems. *Computer Music Journal* 24(1): 57-69, 2000.

[5] S. Fels. Intimacy and Embodiment: implications for art and technology. *ACM Multimedia*, 13-16, 2000.

[6] N. Alm, S. Abe and N. Kuwahara. International Workshop on Cognitive Prostheses and Assisted Communication. *Intelligent User Interfaces*, 14, 2006.

[7] K. Kuwabara, N. Kuwahara, S. Abe and K. Yasuda. Using Semantic Web Technologies for Cognitive Prostheses in Networked Interaction Therapy. *Intelligent User Interfaces Workshop on Cognitive Prostheses and Assisted Communications*, 1-5, 2006.

[8] A. Utsumi, D. Kanbara, S. Kawato, S. Abe and H. Yamauchi. Vision-based Behavior Detection for Monitoring and Assisting Memory-Impaired People. *Intelligent User Interfaces Workshop on Cognitive Prostheses and Assisted Communications*, 10-15, 2006.

[9] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. Embodied conversational agents. MIT Press, 2000.

[10] J. N. Bailenson and N. Yee. Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. *Psychological Science* 16: 814-819, 2005.

[11] K. Perlin. Real Time Responsive Animation with Personality. *IEEE Trans. Visualization and Computer Graphics* 1(1): 5-15, 1995.

[12] M. Funk, K. Kuwabara and M. J. Lyons. Sonification of Facial Actions for Musical Expression. *International Conference on New Interfaces for Musical Expression*, 127-131, 2005.

[13] L. Barrington, M. J. Lyons, D. Diegmann and S. Abe. Ambient Display using Musical Effects. *Intelligent User Interfaces*, 372-374, 2006.
[14] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. *IEEE Computer Vision and Pattern Recognition*, 511-518, 2001.
[15] C. R. Wren, A. Azarbayejani, T. Darrell and A. P. Pentland. Pfinder: Real-Time Tracking of the Human Body. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19(7): 780-785, 1997.
[16] M. Turk. Visual Interaction With Lifelike Character. *IEEE Conference on Automatic Face and Gesture Recognition*, 368-373, 1996.
[17] R. S. Sutton and A. G. Barto. Reinforcement Learning. MIT Press, 1998

List of authors