iMime: An Interactive Character Animation System for use in Dementia Care

Andreas Wiratanaya[1] *Michael J. Lyons*[1] *Nicholas J. Butko*[2] *Shinji Abe*[1]

[1]ATR IRC Labs
2-2-2 Hikaridai, Seika-cho, Souraku-gun, Kyoto
{wiratanaya, michael.lyons}@gmail.com

[2]Department of Cognitive Science
UC San Diego
nbutko@cogsci.ucsd.edu

ABSTRACT

We describe the design and implementation of an interactive character animation interface. The system analyzes the attentive state and aspects of the affective behaviour of a viewer using input from a video camera and uses this to control the behaviour of a cartoon-like animated character. Using the interaction metaphor of a mime artist, we design the system to encourage viewer attention and interaction, with adaptation using an online reinforcement learning based on the viewer's attentive state. This work is ultimately aimed at developing a system to support the care of dementia sufferers.

Author Keywords

Dementia care; gestural interaction; attentive interfaces

ACM Classification Keywords

H.1.2 User/Machine Systems, H.5.2 User Interfaces

INTRODUCTION

The development of prostheses for impaired cognitive abilities is an important and active topic of current research [1,2]. In the case of dementia treatment the requirement for constant care and attention can create a burden for the patient's family members [1]. The resulting stress can have a negative effect on the patient's well-being. One way to reduce this stress is to entertain the patient with audio-visual media which capture and hold the attention over a period of time providing at the same time some relief of the burden of care for the caregiver.

Recently Kuwabara *et al*. [1] 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 memory stimulating images from their past. To add user interactivity to this application, Utsumi *et al.* explored a content switching strategy [2] for attracting and maintaining the attention of video watchers. This uses the patient's gaze direction as a measure of attention, switching to a different channel whenever the patient starts to lose interest.

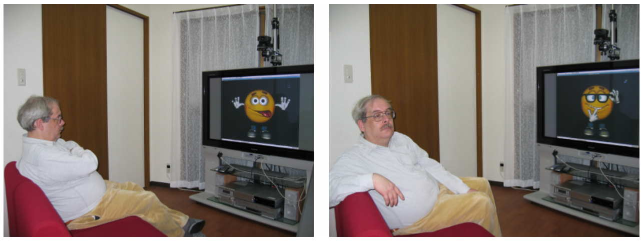

Figure 1. iMime prototype: user interaction takes place non-verbally via input from video cameras.

Here, we describe the design and implementation of a novel interactive interface which substantially extends the concept of maintaining the patient's interest though response-dependent adaptation of content display. Instead of reminiscence video contents we use a much more interactive content in the form of a real-time animated character. The simulation of entertaining, animated characters is an active field of research in computer graphics, however to the best of our knowledge this is the first application of this technique to dementia care. We present a framework for the interaction between a human and a virtual character targeted at the special requirements in the treatment of dementia patients. As middle to late stage dementia patients often suffer from a severely impaired capacity for verbal communication we base our system primarily on visual, non-verbal interaction. This led us to consider the metaphor of the non-verbal performance of a *mime*. As can be quite commonly observed in street performances, mimes are masters of non-verbal communication. Our system draws inspiration from the ability of skilled mime performers to attract and hold attention and entertain, using non-verbal interactive behaviour.

IUI'07, January 28–31, 2007, Honolulu, Hawaii, USA.
Copyright 2007 ACM 1-59593-481-2/07/0001...$5.00.

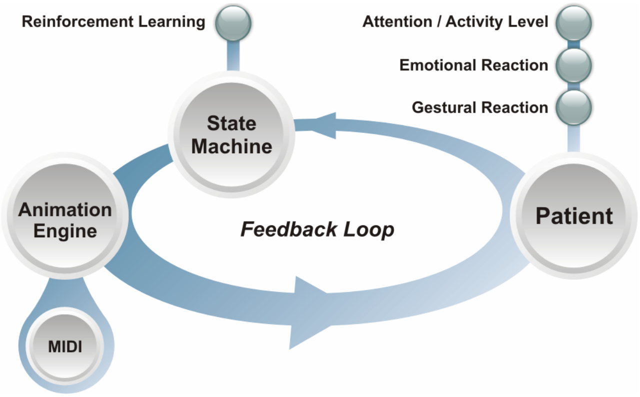

Figure 2. Schematic of the iMime system.

DESIGN

A general schematic of the system is shown in Figure 2. It illustrates the interaction flow between the user (patient) and virtual character generated by the animation engine. We designed our system with the metaphor of a street mime performance in mind. In a street mime performance the actor typically makes a few stylized or humorous actions then freezes. To prompt further action an observer usually has to take some form of action: either donating some money or reacting in a certain way to the mime's behaviour. In this fashion a mime is able to bootstrap interaction with strangers without relying on verbal explanation.

We implemented this interaction metaphor as follows: the viewer's movement is recorded by multiple video cameras at different scales. Computer vision algorithms are used to analyze the appearance and movement of the patient 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 into showing some reaction which in turn serves as an input to the vision system, hence completing the interaction loop.

Sensory Input

While a variety of sensors can be used to capture information about the attentional state of user, the intended use for the system in dementia care restricts the type of input sources to non-verbal and non-intrusive communication channels. In our prototype we use two cameras which capture the patient at different scales: one camera is focused on the face while the other one captures the entire upper body.

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. An observing user will merely look at the presented animation, whereas an interested user may show unconscious facial reactions. An engaged user may additionally respond with full body gestures. The vision system implemented is capable of:

- determining whether or not the user is looking at the display
- recognizing the current head orientation
- classifying the overall body motion
- recognizing emotion primitives such as smile or frown
- recognizing different basic gestures

Adapting to the Patient

A general problem in scripted animation systems or video contents is repetitiveness. Even the best scripts become boring after some time if the user can observe 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 which controls the animated character with an online reinforcement learning system. This system analyzes the attentional reaction of the patient and devises a strategy to maximize the user's attention. Instead of greedily choosing the locally best solution the system is able to make a choice which could lead to a better solution in the future even if this includes taking a locally sub-optimal path.

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 prototype which allows controlling the facial expression of the animated character by playing MIDI files.

IMPLEMENTATION

Animation Engine

A mime conveys information exclusively over two channels: facial expression and body language. This has to be kept in mind when choosing a model for the animated character. From consultation with a clinician involved in research on dementia care, we learned that overly realistic human models would most likely not be accepted. Animal models on the other hand while being cute impose limits on the usable range of body language. Finally, we chose the relatively abstract model shown in Figure 3 for its high expressiveness, flexibility, and cartoon-like character.

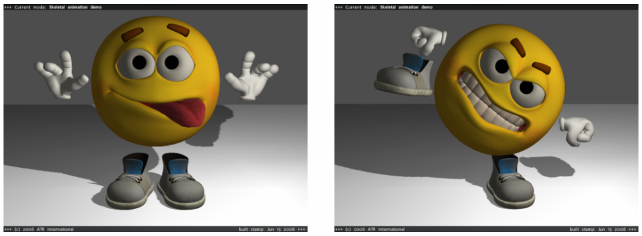

Figure 3. Animated character used by the iMime system.

Facial Animation

Facial animation is a well-studied problem in computer graphics. We implemented a parameterized muscle model but found that the results were not expressive enough for our purpose while at the same time varying muscle parameters proved very unintuitive. We therefore decided to use handcrafted morph targets. This allows decomposing facial expressions into different primitives such as raising an eyebrow, different mouth shapes or tongue movement. Some examples are shown in Figure 4. Primitives can be blended together linearly to form composite expressions of high variability. The currently used model has 42 different primitives and core expressions thus forming a 42-dimensional expression space.

Body Animation

Movement of the limbs typically includes rotations which constitute non-linear motion. It is therefore not possible to define basic body poses on vertex basis and blend them linearly as done with the facial expressions. However in order to be able to arbitrarily combine movement primitives to create more variation we have exactly this requirement. The standard solution to this problem is to use a skeletal animation approach. The model is augmented with a skeleton consisting of different bones arranged in a hierarchical fashion inspired by the human body. For example moving the upper arm will also move the lower arm which in turn moves the hand. Every bone is assigned a region of influence on the model. The skeleton used in our prototype is shown in Figure 4. Animations can now be parameterized and blended easily using rotation angles for each bone. Also predefined animations can be given a more lifelike appearance by convolving the parameters for each joint with a Perlin noise function.

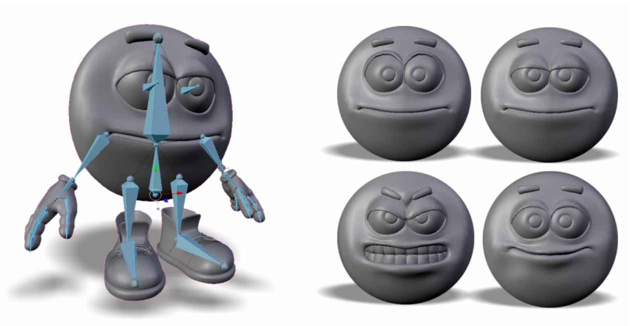

Figure 4. Structural components used by the animation engine. Left: the skeletal system used for animating the body. Right: several facial expression morph targets.

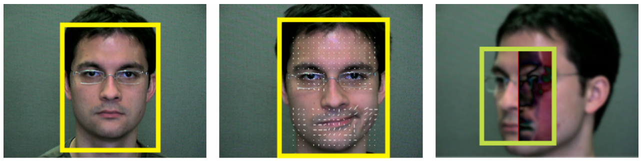

Figure 5. Interpreting facial cues. Left: face position detection. Center: classifying non-rigid motion using optical flow. Right: attention classification using the symmetric difference image.

Vision System

Our prototype uses two cameras to capture the patient at different scales. The camera focused on the face is used to extract emotional and attentional information while the camera focused on the upper body extracts gestural information.

Analyzing the face

We extend a system previously described by our group [3,4] to meet the requirements for this project. The system uses combined automatic face detection and optical flow to classify facial expressions and is able to discriminate between rigid and non-rigid movement of the head.

In order 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 L_2-distance between the left and right face half after applying a median filter to the image as shown in Figure 5. We can also determine whether the patient is looking to the left or to the right with a similar method: we shrink the face rectangle to the skin colored area of the face and compute the center of gravity of all significant edges found in this area. If the head is turned to the left, the face rectangle moves to include the side profile of the head on the left side while the other half covers almost exclusively skin colored area. We can therefore observe strong edge dominance in the left half of the face rectangle whereas almost no significant edges can be found in the right half. Combining the two aforementioned techniques allows drawing sufficient conclusion about the head orientation. While the first algorithm assumes more or less uniform lighting of the face, the second algorithm is robust against lighting variations.

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

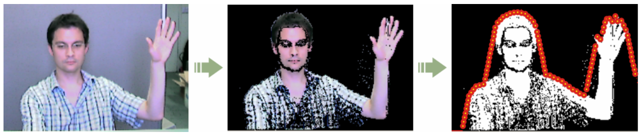

Figure 6. Gesture analysis. Left: input image. Center: background removal. Right: draping the outline.

Analyzing the upper body

To classify the patient's pose we first separate the foreground pixels (patient) from the background pixels. Since the envisioned application of the system is in an indoor environment we can use a Gaussian model [5] to describe the background, in which every pixel is associated with a full covariance matrix. Figure 6 shows the typical results after thresholding the input image using the Mahalanobis distance. The image is then binarized and a mass-spring model [6] 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 is shown in the right image of Figure 6. The algorithm returns a vector of height values which are normalized and correlated to previously acquired reference poses for classification.

Reinforcement Learning

We implemented a real-time, online reinforcement learning system [7] which consists of two components: a *model estimator* for computing 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, with a reward structure based on information about the interest level of the patient provided by the vision system.

SYSTEM INTEGRATION

In order to get a first impression of the possible interactions, we have integrated all components into a prototype system and defined some basic behaviours using the available sensor data. The animated character is aware of user presence. If no user is seen he 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. Once a user 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 properly capture the face. During interaction with the user the character will mimic facial expressions such as smile or eyebrow movement or gestures such as waving. If the user does not interact for a certain amount of time the character will first visibly ponder, then point at him and show one example gesture followed by a rewarding animation in case it is copied correctly by the user or a no-no gesture in case of failure. If the user tries to confuse the system by performing erratic movements, the character will stop what he is doing, look at the user puzzled and scratch his head.

CONCLUSIONS AND FUTURE WORK

This paper descrbied the design and implementation of a character animation system intended to attract and entertain 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 field tests with dementia sufferers. Our ultimate aim is to refine the system to the point where it can become an effective tool in dementia therapy, reducing the suffering of the patient as well as easing the burden on the caregiver.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institute of Information and Communications Technology and the NSF-JSPS EAPSI program.

REFERENCES

1. Kuwabara, K., Kuwahara, N., Abe, S. and Yasuda, K. Using Semantic Web Technologies for Cognitive Prostheses in Networked Interaction Therapy. *Proc. Workshop on Cognitive Prostheses and Assisted Communications, IUI 2006*, 1-5
2. Utsumi, A., Kanbara, D., Kawato, S., Abe, S. and Yamauchi, H. Vision-based Behavior Detection for Monitoring and Assisting Memory-Impaired People. *Proc. Workshop on Cognitive Prostheses and Assisted Communications, IUI 2006*, 10-15
3. Funk, M., Kuwabara, K., and Lyons, M.J. Sonification of Facial Actions for Musical Expression. *Proc. NIME-05*, 2005, 127-131
4. Barrington, L., Lyons, M.J., Diegmann, D., and Abe, S. Ambient Display using Musical Effects. *Proc. IUI'06*, 2006, 372-374
5. Wren, C.R., Azarbayejani, A., Darrell, T. and Pentland, A.P. Pfinder: Real-Time Tracking of the Human Body. *IEEE PAMI*, 1997, 780-785
6. Turk, M. Visual Interaction With Lifelike Characters. *Proc. 2nd Conf. on Automatic Face and Gesture Recognition*, IEEE (1996)
7. Sutton, R. S. and Barto, A. G. Reinforcement Learning. *MIT Press*, Cambridge, MA, 1998