Virtual PAT: A Virtual Personal Aerobics Trainer

James W. Davis Aaron F. Bobick
MIT Media Lab
20 Ames Street, Cambridge, MA 02139
jdavis@media.mit.edu, bobick@media.mit.edu

Abstract

A prototype system for implementing a virtual Personal Aerobics Trainer (PAT) is presented. Unlike workout video tapes or TV exercise shows, this system allows the user to create and personalize an aerobics session to meet the user's needs and desires. Various media technology and computer vision algorithms are used to enhance the interaction of a virtual instructor by enabling it to watch and talk to the user. Throughout the paper, we report the system components and discuss the advantages, problems, and extensions of the design.

1. Introduction

In this paper we discuss the design and implementation of a prototype virtual Personal Aerobics Trainer (PAT). The underlying motivation for building such a system is that many forms of media that *pretend* to be interactive are in fact deaf, dumb, and blind. For example, most of the aerobics workout videos that one can buy or rent present an instructor that blindly expels verbal re-enforcements (e.g. "very good!") whether or not a person is doing the moves (or is even in the room!). There would be a substantial improvement if the TV just knew whether or not a person was moving in front of it. A feeling of awareness would then be associated with the system. And because of the repetitiveness of watching the same exercise videos, this "programmable" system heightens the interest of the user by allowing the design of specialized workouts (e.g. exercising only the upper body).

The system here creates a personalized aerobics session for the user and displays the resulting instruction on a large TV screen. Here the user can select which moves, their duration, the music, and which instructor are desired for the workout. Depending on the mood of the user, the choice of instructor could make a large difference in the workout. For instance, if the user were tired and required strong motivation during the workout, a brash Army Drill Sargent as the instructor might be an appropriate choice. Along those lines, the prototype system here makes available an Army Drill Sargent character as the instructor. The session is scripted by the user, and is then automatically generated and started when the user enters the area in front of the TV screen (See Figure 1).

Figure 1. Virtual PAT. A virtual Personal Aerobics Trainer. Photo credit: Webb Chappell. Copyright: Webb Chappell 1998.

In addition to the video demonstrations, the user periodically receives audio feedback from the virtual instructor on how he/she is currently doing. We place a video camera (or multiple cameras) in the room and use real-time computer vision techniques to recognize the aerobic movements of the user. Using the output of the vision system, the virtual instructor then responds accordingly (e.g "good job!", if the vision system recognizes that the user is performing the aerobic move). This vision technology is different from many other sensing technologies in that the user need not wear

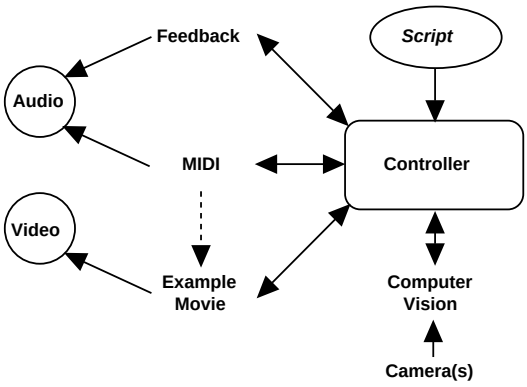

Figure 2. Component diagram showing the interconnection of the media, control, and computer vision modules within the PAT system.

any special clothing/devices or be tethered to machines with bundles of wires. This enables the experience to be more natural and desirable [2, 9]. We now explain the design of the system.

2. System design

The PAT system is a modular design of media and vision components connected to a controller (See Figure 2). All software[1] was written in C++ and run across SGI R10000 O2 computer systems (though we believe the components could be placed within a lower-end hardware setup). For the current system, the output video (a series of movie clips showing the instructor performing the moves) is sent to a large TV screen, the feedback is routed through the audio channel, and the music is stored/played in the form of MIDI music files.

Because the current version of the system uses real video clips it would be tedious to record all possible feedbacks during all possible moves. Therefore, the audio is decoupled from the video (i.e. the lips of the instructor do not move, as if speaking the lines). The separating of audio feedback from the visual example permits the simple combination of various spoken feedback phrases and multiple aerobic examples. It would be possible to integrate the spoken feedback and the visually displayed by using a computer graphics model of the instructor, rather than using movie clips. The correct combined state of the instructor (e.g. doing jumping jacks while complementing the user) would be controlled and rendered at run-time. It might even be quite fun to have virtual cartoonish-like characters as the instructors. Each character could have their own "attitude" and behavior [1], which would possibly increase the entertainment value of the system. But in the current system, we chose to use stored movie clips for simplicity. Also, to date we have only recorded one instructor though the system is designed to have multiple instructors from which the user can choose.

When personalizing (or creating) the aerobics workout session, the user need not worry about specific timing information of the moves, feedback, or music to coordinate the different media events or structure of the session. These components each have underlying temporal information (e.g. movie length, tempo of music) which is communicated between different modules. The user need only manipulate a simple script of the workout session.

We now give a brief explanation of each of the components in the system. The components residing on different machines have a direct communication line to a main control program using the Parallel Virtual Machine (PVM) software designed for heterogeneous network computing.

2.1. Scripting

Since most instructional systems employ some underlying notion of event ordering, we can exploit this to allow the user to create and structure a personalized session. The system was designed so that each session is guided from a simple script which controls the flow of the session (similar to [6]). Included in the script are the names for the workout moves, the time allotted for each move, the choice of music (song titles) for the workout, and the instructor (by name) to run the session. This allows the user to easily choose their own tailored workout. Placing the available options and examples in a GUI would make the script creation process quite trivial, but for now it is created in a text editor. The controller loads this script and initiates the program. The program is immediately available upon instantiation of the system with a script (it is not generated or compiled off-line).

2.2. Controller

A simple state-based controller was developed to run the workout script and act as a central node connecting the various modules. The controller consists of seven states: Pause, Startup, Introduction, Workout, Close, Shutdown, and PreClose. The system begins in the Pause state, where it resides until a person enters the space, determined by querying the vision module. The entering person event triggers Startup state which

[1] All media components were developed using SGI's Digital Media utilities/libraries.

Figure 3. Video output of the virtual instructor.

opens the video window and performs some system preparation. Next is the Introduction state, where the virtual instructor appears and welcomes the user. After the brief introduction, the system loops in the Workout state (a loop for each move in the session) until all moves are completed. Feedback comments from the instructor are given here (comment categories are determined by the recognition results from the vision module). Then a Close state gives the final goodbye comments from the instructor, followed by the Shutdown state where the display is turned off and then system cleanup is initiated. There is an additional PreClose state which is entered if the user prematurely leaves the space (recognized by the vision module). Here, the instructor realizes the user is no longer present, and then starts a shutdown (or pause) of the system (the program will not continue if no one is there to participate).

2.3. Instruction by example

In the TV display is a window showing the instructor performing the aerobic moves. Figure 3 shows a few key frames from some stretches and aerobic exercises. This is provided so the user can follow along, watching how to perform the moves, as done with a real instructor.

Currently, a set of movie clips, showing a full view of the instructor, is used. In each clip (for each move), the instructor performs a single cycle of the move. This clip is looped (or swinged) for the duration of the time devoted for that move. Given the beat track of the music, the speed of the current movie clip can be altered to be in synchronization with the music currently being played. (It can be quite annoying for the video/user to be moving out of synchronization with the music if there is a strong beat.) We now discuss a music format that can be used to synchronize the example movies with the music.

2.4. MIDI music and synchronization

Music is used in aerobics primarily to regulate the speed and consistency of the moves as well as to maintain the energy of the exerciser. We currently use MIDI music files/songs as our source of music: a MIDI file is a symbolic description of the music similar to the sheet music or score for a song. By connecting a synthesizer and speakers to a computer (some computers provide an internal software synthesizer), the song can be played by sending the MIDI file from the computer to the synthesizer which then generates the music. One advantage of using MIDI is that it is easy to alter the music by simply inserting, deleting, or changing individual MIDI events. Another important feature of MIDI is that there is a tempo associated with each song.

To suit our purposes, we would like to synchronize our instructor's movements from the movie clips to the beat of the music. Using the MIDI format, we can add a silent beat track — a metronome click — using the tempo of the song. The beats act as a trigger to which the instructor movies can be synchronized. The movie cycle duration is set to be equal to the time required for a fixed number of beats, and the movie is looped/swinged after a certain beat count.

The beat track is added as an additional track to the MIDI file and is set at a non-perceivable volume. This way, the computer can watch for the beat events while not actually playing audible notes. We note that there exists methods for accomplishing the above beat track synchronization using other musical media forms (e.g. a CD); such a method is found in [8].

At times, there can be a bit of a jump when looping the movie clips in synchronization with the beat track of the music. The actual time *measured* between beats in a song and also the response time of the movie adjustment can sometimes vary slightly. This causes the looping to become minutely jagged at the transition for the next loop. More precise control and reading of the beat track can fix this problem. Another synchronization problem can arise if there is substantial video buffering (or lag) from the video generator to the display unit, causing the virtual instructor to appear as dragging behind the beat. Due to the above problems, we built two systems where one version synchronizes the video and music and another version where the two

are unsynchronized (the music was carefully selected in this case).

The MIDI songs for the workout are all queued up at the beginning of the session so one song plays immediately after another. At the end of the session, the music is stopped.

2.5. Feedback on performance

Each time a new aerobic move begins, the feedback module is activated and a brief statement about the new move is given. For some moves, the comment may give the name of the movement (e.g. "It's time for some jumping jacks") or for other moves explain their purpose (e.g. "This move is going to work the shoulders"). For the feedback to the user the system contains many *positive* comments (e.g. "good job!", "fantastic!") and many *negative* feedback comments (e.g. "get moving!", "concentrate!"). Whenever the controller decides it is time for a feedback comment[2], it randomly picks a comment from the appropriate category. This way, one does not always here the same comment many times in a row or hear the same ordering of comments. There is an opportunity here to record very expressive comments for the system, which increases the entertainment value of the system as well as its usefulness.

3. Recognizing aerobic movements with computer vision

In this section, we give a brief explanation of the methods used for recognizing the movements of the user. We use real-time computer vision techniques to "watch" the user and determine if he/she is currently performing the same move as the instructor. A background-subtracted image of the room yielding only a silhouette of the person [4, 2, 9] is used as the input to the vision system.

3.1. Robust silhouette extraction technique

The silhouetting method used in this system is based on the optical blocking (or eclipsing) of specialized non-visible light (e.g. infrared light) rather than the color differences between the person and background. In our approach, the person is always able to wear any color clothing, and the end effect is the same (unlike the color-based methods). The method also permits the display of video graphics behind the user on the background. Figure 4 shows the environment configuration. This system is similar in nature to the infrared designs of [7, 5].

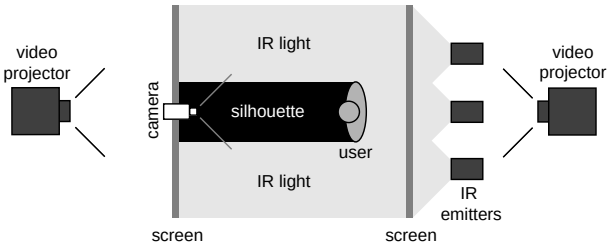

Figure 4. Conceptual drawing of blocking (eclipsing) infrared light from the camera to generate a silhouette of the person.

We can see the silhouetting process in Figure 5. Figure 5(a) shows a standard camera view of someone standing in front of the back-wall projection screen with graphics displayed. In Figure 5(b), we see the same picture but now with infrared light directed through the back screen using six infrared light (840 nm) emitters[3]. By placing an infrared-pass/visible-block filter over the video camera lens, a brightly lit screen (with no graphics visible) and a clean silhouette of the user is seen, as shown in Figure 5(c). The infrared light is not visible to the human visual system and thus one sees only video projected on the display screen (as shown in Figure 5(a)). We point the reader to [4] for further details on the approach. This silhouette form is then further processed to recognize the movements of the user.

3.2. Motion Templates

Recently, we have developed real-time computer vision methods for recognizing large-scale body movements such as aerobic exercises. That work constructs temporally-collapsed motion templates of the participant's silhouette, and measures shape properties of that template to recognize various aerobic exercise (and other) movements in real-time.

To show an example of such motion templates, Figure 6 presents templates generated from the infrared silhouettes for the movements of left-arm-raise (left-side stretch) and fan-up-both-arms (deep-breathing ex-

[2]The system checks every movement cycle to see if the user is complying. A *negative* comment is given every cycle until the user performs the move. If the user is performing the move, a *positive* comment is given approximately every few cycles or some predetermined wait period (whichever is smaller).

[3]The camera has no infrared-blocking filter and is thus sensitive to infrared light.

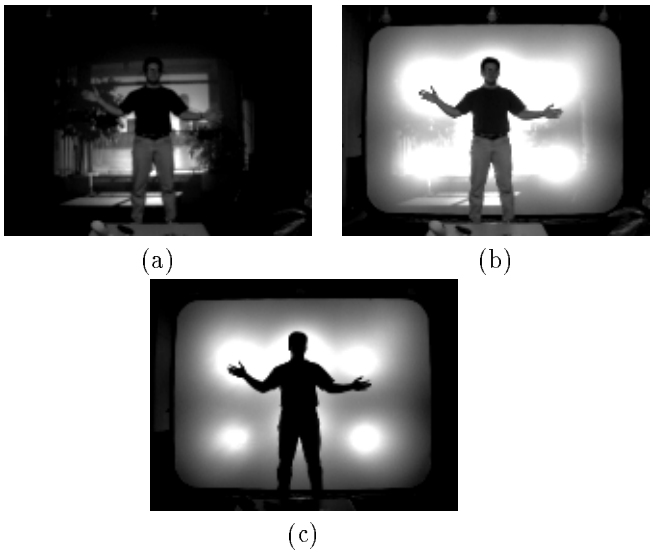

Figure 5. Infrared light. (a) The person in front of a video projection screen. (b) Now with infrared light directed through the screen. (c) The image filtered through a visible-block filter.

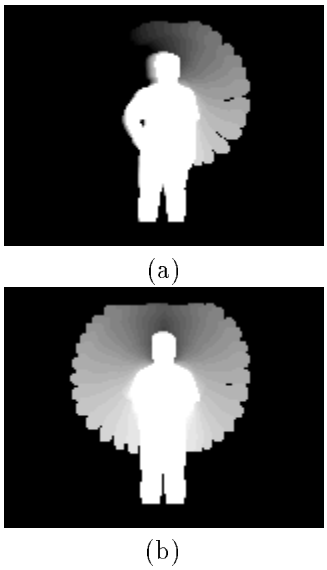

Figure 6. Example motion templates for infrared silhouettes. (a) Motion template for left-arm-raise (left-side stretch). (b) Motion template for fan-up-both-arms (deep-breathing exercise stretch).

ercise stretch). The motion of the user is encoded in the varying graylevels within the template.

For recognition of these moves, statistical pattern recognition techniques are applied to moment-based feature descriptors of the templates. The system employs user training to get a measure of the variation which arise from different people. This approach easily extends to multiple camera views of the person; see [3] for details on the algorithm.

4. Future possibilities

Presently, the system is designed for a single user. To increase the usefulness of the system, it could be designed to run with multiple people (as in a class). We could accomplish this by spreading out the people inside the room so the cameras could get an unobstructed view of each person. The system could also periodically switch its focus from one person to the next (by toggling camera views) to save in processing. However, better vision techniques that allow occlusions from multiple people would be desirable. Also having the system understand and reason about other objects (e.g. weights, towels) would increase its awareness and expressiveness of its environment.

To extend the sensing and robustness of the system as a whole, we could incorporate other types of sensors. For instance, we could run a trigger wire across the top of an exercise step (used for step aerobics). Then it would be easy to verify whether the person was in fact on the step. With this, we can also check the periodicity of the stepping and verify synchronization of the user with the instructor. There are also wireless heart-rate monitors available which can transmit their data to remote computers. With this accessory, the user can have his/her heart-rate displayed on the TV screen while progressing through the workout session. A program could be written to check this information and watch for harmful levels of strenuous activity.

The large video screen behind the user, while necessary for the infrared silhouetting procedure, can also be used to enhance and make the experience more rich and immersive to the users. For example, we could add virtual people to the back screen to make the user feel as if taking part in a group workout. There is also an opportunity here for putting entertainment on this screen for benefit of an audience, rather than just the user. For instance, it may be humorous for an audience (or those waiting their turn) to watch the virtual class toying with the user (unbeknownst to the user, of course). If additional video screens were employed, then we could place the user entirely within some virtual world.

Another video window could be shown (adjacent to the instructor window) which contains a live video feed of the user as seen from a camera mounted near the TV screen. This video mirror (with the video stream reflected) would serve the purpose of giving a visual aid to the user involved in the current session. The person can watch themselves and see how they look while performing the moves. We could also augment the video with graphics to better aid in responding to the user (e.g. drawing a straight line down the user's back to emphasis to the user to keep his/her back straight).

Lastly, a record of past sessions for an individual could be kept (e.g. retaining the moves the person performed, and how well he/she executed those moves). With this past information available, the system could be designed to respond with more informed comments like, "Your doing much better today!". This capability would move the PAT system into the highly personalized interaction category.

5. Conclusion

In this paper, we described a prototype system for a virtual Personal Aerobics Trainer (PAT). Users can select which aerobic moves, which songs, and which instructor they want for their workout session. The resulting program guides the user through the workout while watching and commenting on the user's performance.

The system presents various movie clips of the instructor, which can be ordered and synchronized with the music. A computer vision system watches the user throughout the session, and reports to the virtual instructor how the user is doing. With this information, comments are given by the instructor for pushing or complementing the user.

This system moves beyond the highly non-interactive media forms of video tapes and TV shows by watching and responding to the user. We feel that many future exercise systems will be more interactive and less passive, and perhaps one day will be commonplace within the home environment.

6. Acknowledgments

We first would like to especially thank Andy Lippman for playing out the role of the Army Drill Sargent. His sense of fun made that character come to life. We would also like to thank those people who provided their time sweating for us by training up the system. This work is supported in part by the Digital Life consortium of the MIT Media Laboratory.

References

[1] Blumberg, B. Old tricks, new dogs: ethology and interactive creatures. PhD dissertation, MIT Media Lab, MIT, 1996.

[2] A. Bobick, S. Intille, J. Davis, F. Baird, L. Cambell, Y. Ivanov, C. Pinhanez, A. Schutte, and A. Wilson. The KidsRoom: Action recognition in an interactive story environment. *Presence (to appear).*

[3] Davis, J. and A. Bobick. The representation and recognition of human movement using temporal templates. In *Proc. Comp. Vis. and Pattern Rec.*, pages 928–934, June 1997.

[4] Davis, J. and A. Bobick. A robust human-silhouette extraction technique for interactive virtual environments. In *IFIP Workshop on Modelling and Motion Capture Techniques for Virtual Environments*, Geneva, Switzerland, Nov. 1998.

[5] Ishii, H., and B. Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. In *CHI'97*, pages 234 – 241, 1997.

[6] C. S. Pinhanez, K. Mase, and A. F. Bobick. Interval scripts: A design paradigm for story-based interactive systems. In *CHI'97*, pages 287–294, Atlanta, Georgia, Mar. 1997.

[7] Rekimoto, J. and N. Matsushita. Perceptual surfaces: towards a human and object sensitive interactive display. In *Workshop on Perceptual User Interfaces (PUI-97)*, pages 30 – 32, October 1997.

[8] Scheirer, E. Tempo and beat analysis of acoustic musical signals. *Acoustic Society of America*, 103(1):588–601, Jan 1998.

[9] Wren, C., Azarbayejani, A., Darrell, T., and A. Pentland. Pfinder: Real-time tracking of the human body. In *SPIE Conference on Integration Issues in Large Commercial Media Delivery Systems*, 1995.