M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 436
Vision and Modeling group

Virtual PAT: A Virtual Personal Aerobics Trainer

James W. Davis **Aaron F. Bobick**
MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139
(jdavis@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 the character by enabling it to watch and talk to the user (instead of just the user watching the TV). 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). This system creates a personalized aerobics session for the user and displays the resulting instruction on a TV screen. Here the user can choose which moves (and for how long), which 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 would make a great choice. Along those lines, the prototype system here makes available an Army Drill Sargent character as the instructor. The session created by the user is then automatically generated and begins when the user enters the area in front of the TV screen (See Figure 1).

In addition, the user periodically receives audio-visual feedback from the virtual instructor on how he/she is currently doing. To accomplish this, we place video 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 instructor then responds accordingly (e.g "good job!" if the vision system recognizes that the user is performing the aerobic move correctly). This vision technology is different from many other sensing technologies in that the user need not wear any special devices or be tethered to machines with bundles of wires. This enables the experience to be more natural and desirable [2, 6, 4].

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, many 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 even is in the room!). There would be a substantial improvement if the TV just knew whether or not a person was moving in front of the TV. 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 specialized workouts (e.g. exercising only the upper body). We now explain the design of this system.

Figure 1: PAT session. The user watches the TV screen while following the virtual instructor (an Army Drill Sargent).

2 System design

The PAT system is a modular design of media and vision components [3] (See Figure 2). All software[1] was written in C++ and run on SGI R10000 O2 computer systems (though we believe all the components could be placed within a much smaller hardware setup). The output graphics of the system is sent to the TV screen, and is shown in Figure 3. The display is composed of three sections/windows: one for the instructor performing the moves, one for the feedback from the instructor, and one for a video mirror showing the user in the room.

We chose to decouple the feedback from the movement examples to eliminate the complexity of stored sequences. It would be very hard to gather video

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

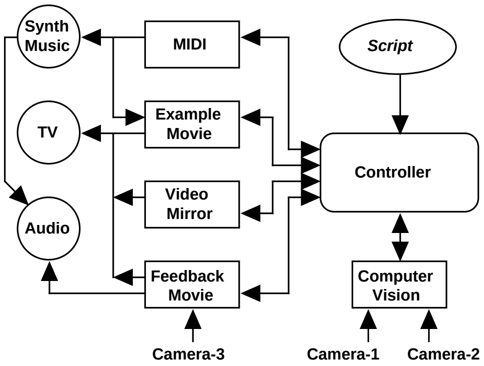

Figure 2: Component diagram showing the interconnection of the media, control, and computer vision modules within the PAT system.

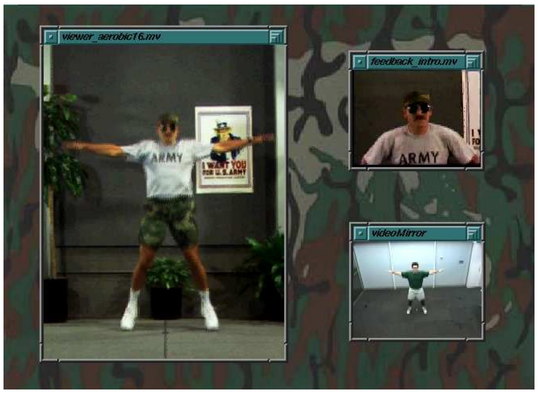

Figure 3: The TV display is composed of three windows. The leftmost window shows the instructor as he performs the moves. The top-right window gives the feedback to the user. The bottom-right window acts as a mirror so the user can also watch themselves as the perform the moves.

clips of all possible feedbacks with all possible aerobic moves. Therefore the splitting of feedback from visual example permits the trivial use of various feedback lines and multiple aerobic examples. It is also possible to collapse the feedback and aerobic example back into one frame by using a computer graphics model of the instructor. Here, the correct state of the instructor (e.g. doing jumping jacks while complementing the user) could 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 this system, we chose to use stored movie clips for simplicity. While we use only one instructor (the Army Drill Sargent), the system is designed to have multiple instructors from which to choose.

As for personalizing (or creating) an aerobics workout session, the system designer (or 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 events each have underlying temporal information (e.g. movie length, tempo of music) which can be communicated between different modules and used without the manual intervention of the designer (i.e. the programmer/user need not know about, or hardcode, the specific timings of the movie lengths or music tempo).

We now give a brief explanation of each of the components in the system. The components have a direct communication line to a main control program using the Parallel Virtual Machine (PVM) software used for heterogeneous network computing[2].

2.1 Music with beat track

Music is used in aerobics primarily to keep the energy flowing and regulate the speed and consistency of the moves. In this system, we currently use MIDI music files/songs as our source of music, where a MIDI file is a symbolic description of the music (i.e. 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/heard by sending the midi file from the computer to the synthesizer which then generates the music. One advantage of using MIDI music is that it is easy to alter the music by simply inserting, deleting, or changing individual MIDI events. Another feature of MIDI is that there is a tempo associated with each song.

To suit our purposes, we need to synchronize our instructor's movements from the video clips to the beat of the music. Using the MIDI format, we can easily add a beat track (similar to a metronome click) to each song and use this as a trigger to which the instructor movies can be synchronized. This 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 the notes.

The MIDI file (augmented with the beat track) is sent from the computer to a synthesizer. In our computer systems, there is an output temporal buffer provided for sending midi events, because events can be "sent" faster than they can be "played". Given this configuration, one can't easily determine when each events is actually played. Since we need to be able to know when each beat is "played", we loop back the MIDI stream from the synthesizer (which does not queue events) into the computer which then monitors for the beats. This gives a fairly precise way of reading which events were just played.

A program was written to read the MIDI events (from the loop around) and search for the beats from the added beat track. The tempo of the beats found then directly affect the speed of the instructor's example movie. We would like to note that there exists methods for accomplishing the above task using other musical media forms (e.g. CD), and such a method can be found in [5]. The MIDI songs are all queued

[2] `http://www.netlib.org/pvm3/book/pvm-book.html`

up at the beginning of the session so one song plays immediately after another. At the end of the session, the music is stopped.

2.2 Instruction by example

In the left portion of the TV display is a window showing the instructor performing the aerobic moves (See Figure 3). 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 for the duration of the time devoted for that move, and the speed of the current movie clip is altered to be in synchronization with the music currently being played. (It would be quite annoying for the user to be moving out of sync with the music.) Given the beat track of the music (as discussed above), we can choose an appropriate speed for the movie clip. Since the movie has an associated duration time, we can see how many beats would trigger during this time. Enforcing that there be an even integer number of beats (more natural for faster movements like jumping jacks), we can adjust the speed of the movie as to cycle on this number of beats.

2.3 Feedback on performance

We created a window in the top-right of the TV display specifically for audio-visual feedback of the instructor (See Figure 3). In this window a close-up view (head and shoulders) of the instructor is displayed. The feedback is composed of a short movies with audio. This window is used by the instructor to introduce each move and also to give comments to the user.

Each time a new aerobic move begins, this window is activated and a brief statement about the new move is given. For some moves, the comment may give the name of the move (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"). As 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 system decides it is time for a feedback comment[3], 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.

2.4 Video mirror

The final window in the lower-right of the TV display is a video mirror (See Figure 3). This window shows a live video feed of the user from a video camera (above the TV screen). To give the effect of a mirror (which is commonly found in aerobic and dance studios), the video images are flipped/mirrored before being displayed on the TV screen. This video mirror servers only the purpose to give 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.

2.5 Recognizing aerobics with computer vision

In this section, we give a light explanation of the methods used for watching and understanding the movements of the user. We use real-time computer vision techniques to "watch" the user and determine if he/she is performing the same move as the instructor.

One additional task of the vision system is to monitor the area and make sure someone is actually present in the space. A background-subtracted image of the room yielding only a silhouette of the person [2, 6] is used to track the location of the person. The tracking procedure simply follows the centroid of the silhouette and determines the location, within the video image, of the person. The PAT system starts-up when a person stands in front of the TV screen for a few seconds (passers-by do not trigger the system startup). Also, if the person prematurely leaves the area during the workout session, the system recognizes that the person has left and correspondingly closes down the session.

Recent work in computer vision has show promising results in recognizing large-scale aerobic movements [3]. This research uses temporally-collapsed motion templates to recognize various aerobic exercise (and other) movements in real-time. To ease in discussion here, we point the reader to [3] for details on the algorthim.

We currently have six stretching and aerobic moves entered into the system. For recognition of these moves from the computer vision module, the system was trained using six people performing multiple repetitions of the six moves. This gave a measure of variation which may be seen from different people (though still a small training set) to the vision system. Statistical pattern recognition techniques are then employed for the recognition task.

The current system uses two cameras (approximately 45 degrees apart) aimed at the user. Given the set of moves that the system is to recognize, we can manually determine which camera view best shows the person doing that move in terms of the vision system. Some moves have frontal motion, while other may have large lateral motion; a particular camera position may have a better view than another camera of a move. Therefore, the vision system toggles between the different camera views depending on which move it is currently trying to recognize. There is an area of interesting work yet to be done on *automatically* deciding which views best represent the different moves.

[3] 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 three cycles.

2.6 Scripting

Since most instructional systems employ some underlying notion of event ordering, we can use this to allow the user to create and structure a personalized session. The system was designed so that each session is guided from a script which controls the flow of the session (as in [4]). Included in the script are the names for the workout moves, the time allotted for each move, the choice of music for the workout, and lastly the instructor 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 session creation process quite trivial. The system loads the script and initiates the program. Thus, the program is instantly available upon instantiation of the system with a script, and not generated off-line.

2.7 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 7 states: Pause, Startup, Introduction, Workout, Closing, Shutdown, and PreClosing. The system begins in Pause mode, where it resides until a person enters the space for a period of time. Then begins the Startup mode which opens windows and performs some system preparation. Next is the Introduction state, where a welcome and introduction is given by the instructor. After the brief introduction, the system loops in the Workout state (a loop for each move in the session) until all moves are completed. Then a Closing mode gives the final goodbye comments, followed by the Shutdown mode where the display is turned off and then system cleanup is initiated. There is an additional PreClosing state which is entered if the user prematurely leaves the space. Here, the instructor realizes the user is no longer there, and then starts a shutdown of the system (the program will not continue if no one is there to participate). As previously stated, no hardcoding of media events is necessary, which makes this controller design much less complicated and easy to develop.

3 Problems and extensions

3.1 Current difficulties

With the current vision system (as described above), it is difficult to verify whether the person is performing the aerobic moves in tight synchronization with the instructor (or sometimes even if they are doing the move). It is not easy to pinpoint the exact time for which the move was completed, therefore we cannot robustly check the periodicities of the instructor and user, and look for synchronization. Other techniques and sensors can help alleviate this problem, though.

In our current setup, we export the graphics to a large four-panel display and the audio to a mixer. There is a noticeable (though slight) lag in the resulting TV graphics when compared to the audio stream. This is believed to be caused by some video buffering along that pipeline. A more direct system (unbuffered video streaming) would be more appropriate, and would remove this lag.

At times, there is 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 using the above method can sometimes be a bit unsteady. This causes the looping to become jagged at the transition for the next loop. More precise control and reading of the beat track can fix this problem.

3.2 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 un-obstructed views of each person. Then the cameras could periodically switch their focus from one person to the next. Better vision techniques which allow occlusions from multiple people would be more desirable, though. Also having the system understand and reason about other objects (e.g. weights, towels) would increase its awareness and expressiveness of its environment.

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

To extend the sensing and robustness of the system as a whole, we could incorporate other 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 check whether the person was in fact on the step. With this, we can also check the periodicity of the movement and verify synchronization of the user to the instructor. There are also wireless heart-rate transmitters available which can send their data to remote computers. With this accessory, the user can have their 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 activity.

4 Summary

In this paper, we described a prototype system for a virtual Personal Aerobic Trainer (PAT). The user 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 contains various movie clips of the instructor, which are then ordered and synchronized with the music. A computer vision system using multiple cameras watches the user throughout the session, and reports to the virtual instructor how the user is doing. With this information, comments are given from the instructor to the user (for pushing or complementing them).

This system moves beyond the highly un-interactive media forms of video tapes and TV shows by having the system watch the user (instead of just the user watching the TV). We feel that many future systems will be more interactive and less passive, and that perhaps one day will be commonplace within the home environment.

5 Acknowledgements

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.

References

[1] Blumberg, B. Old tricks, new dogs: ethology and interactive cratures. 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. MIT Media Lab Perceptual Computing Group Technical Report No. 398, MIT, Dec. 1996.

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

[5] Scheirer, E. Tempo and beat analysis of acoustic musical signals. *Acoustic Society of America (in press)*, 1997.

[6] 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.