A Robust Human-Silhouette Extraction Technique for Interactive Virtual Environments

James W. Davis and Aaron F. Bobick

MIT Media Lab, Cambridge MA 02139, USA
{jdavis, bobick}@media.mit.edu

Abstract. In this paper, we present a method for robustly extracting the silhouette form of the participant within an interactive environment. The approach overcomes the inherent problems associated with traditional chroma-keying, background subtraction, and rear-light projection methods. We employ a specialized infrared system while not making the underlying technology apparent to those interacting within the environment. The design also enables multiple video projection screens to be placed around the user. As an example use of this technology, we present an interactive virtual aerobics system. Our current implementation is a portable system which can act as a re-usable infrastructure for many interactive projects.

1 Introduction

When designing interactive environments, it's imperative for the system to be engaging as well as be reliably "aware" of the person (or people) interacting within the space. Many installations are designed with a single large video display, which is the main focus of attention for the user [14, 11, 13, 3]. As for sensing the person in the space, some installations use specialized light, lasers, electromagnetics, or electric field sensing to detect bodies, hands, or objects [14, 11, 13, 8]. Other approaches use similar variants of chroma-keying (i.e. blue-screening) [3], background subtraction [15, 7], or rear-light projection [9] to enable a video camera to extract a silhouette the person, where the person may or may not be required to wear special clothing. We are interested in the latter approaches where a full-body silhouette is visually extracted from the participant; properties of the silhouette such as position, shape, and motion are then used as input for driving the interaction.

The main problem with the current technology for extracting the silhouette is that it relies primarily on the color components of the video signal to perform the extraction. For example, chroma-keying methods require the person stand in front of a background consisting of a uniform-colored wall or screen. The system examines the incoming video signal for the background color of the wall. As long as the person doesn't have that background color on their clothing, the system can extract the person from the video by detecting and removing all the background color in the image. This type of system, commonly used by meteorologists in TV studios, restricts the user not to have the color of the

background anywhere on his/her clothing. If the space is to be used as an interactive environment, the color-based methods as well as the rear-light approach are perceptually obtrusive distracting the user from the interactive experience. Immersion is a strong requirement when building interactive environments [10], and part of this illusion may be broken if such walls are incorporated.

One slight variant of the chroma-keying method is commonly referred to as background subtraction (as used in [3, 7, 15]). A snapshot of the environment containing no people is stored as a reference image. Frames of incoming video imagery are compared with the reference image. Anything that differs, at a pixel-wise level, is assumed to belong to an object (e.g. a person) in the environment. Using the snapshot allows a more natural scene, rather than just a colored or rear-light wall, to be the background. Now however, the colors on the person still need to be *different* everywhere than those of the wall and/or objects behind them. Furthermore, the lighting in the environment must remain relatively constant, even as the person moves about. When these constraints hold, the system works quite well. But if regions of the environment look similar to the person (e.g. a white patch of wall behind a person wearing a white shirt), or if inter-reflection from the person's clothing onto the background is significant, then the person extraction will have either holes or false appendages, respectively. To overcome these problems, a strong model of the body would seem to be required.

In this paper we present a method to overcome the inherent problems associated with the above methods, while opening a new venue for multi-screen interaction. We begin by presenting our design specification for the proposed environment (Sect. 2). Next, we briefly present a virtual aerobics application which uses the proposed environment (Sect. 3), and lastly conclude with a brief summary of the framework (Sect. 4).

2 Design Specification

We divide the specification of the system into three main components. First, the environment itself is examined. We next present how specialized non-visible lighting can be used to enable robust sensing of the participant. Lastly, simple image processing techniques are shown for extracting the silhouette from the video stream.

2.1 The Environment

The prototype environment for showing the utility of the approach consists of two large video projection screens, with one behind and one in front of the user (see Fig. 1). The primary interaction screen is the frontal video display, though video or graphics can be displayed on both screens, enabling virtual objects or interactions on either of these displays. The use of back-projected video screens is necessary (at least for the back wall) for the method which extracts the silhouette of the user in the space (to be discussed). We employ

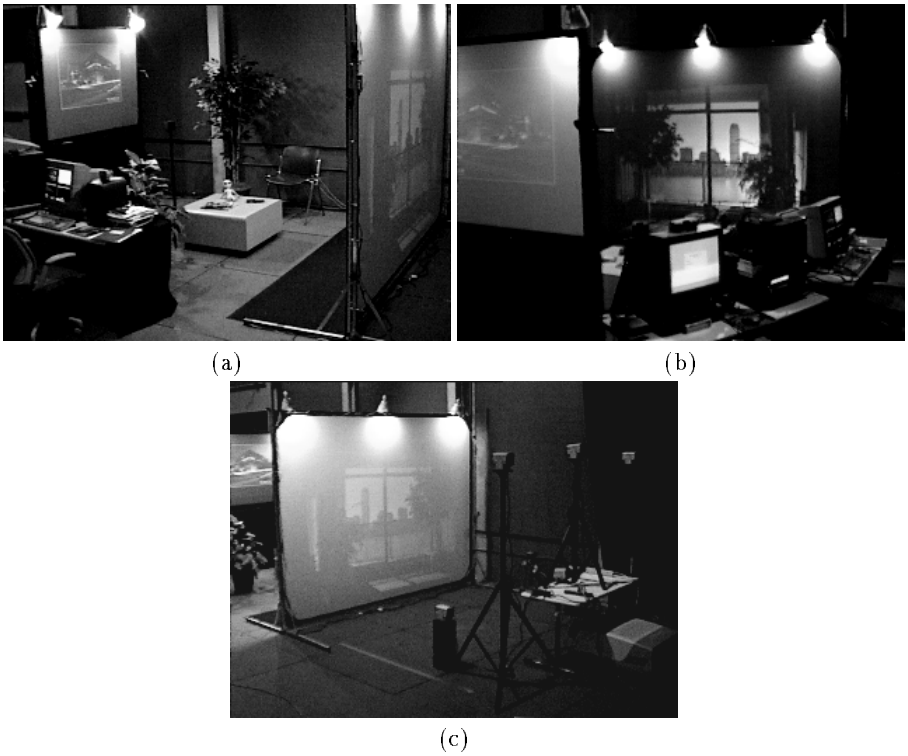


Fig. 1. Dual-screen environment. The environment consists of two video projection screens, with one in front of and one behind the user. (a) View from the outside the rear. (b) View from outside the front. (c) View from behind the environment showing the infrared emitters aimed at the rear screen.

collapsible "snap-on" screens that allow the system to be easily transported to different locations, unlike systems that use large projection TVs.

Behind the user is a 10x8 foot back-projection screen used as the back wall. In front of the user is an 5x4 foot back-projection screen, which is elevated 3 feet off the floor (using two support legs), resembling a large-screen TV. (Later we explain why an elevated smaller screen is used as the front screen instead of a full-sized screen.) The distance between these two screens is 10 feet, large enough not to crowd the user in the space between the screens. Also, the resolution of the projected video on the front screen dictates this pleasing viewing distance.

2.2 Infrared Lighting

To allow the reliable extraction of the frontal silhouette of the user with a "live" video screen behind him/her, we direct invisible infrared light (using 6 consumer 840nm IR emitters) through the large back-projected screen behind the user (see

Fig. 1(c)). These emitters are positioned such that the IR light is distributed across the back screen. By using an infrared-pass/visible-block filter tuned to this wavelength of infrared light, we can then restrict an inexpensive black-and-white video camera[1] placed well in front of the user to see only this infrared light. A person standing in front of the rear screen physically blocks the infrared light diffused through the screen, causing a video camera placed in front of the user to see a bright image with a black silhouette of the person. To get the most flat, frontal view of the person, the camera needs to reside approximately hip-level to the user[2]. Because the camera cannot sit behind the front screen (with the screen blocking the view, and the camera causing shadows on the screen from the projector light) or in front of the screen (such a visible sensor reduces the sense of immersion), we attached the camera to the bottom of an elevated front screen (3 feet off the ground). The elevated front screen resembles a large-screen TV at eye-level in the space and provides an adequate "virtual window" for interactive applications.

One advantage of using infrared light is that many video projectors emit very little infrared light or can be outfitted to do so with infrared-block filters. Therefore, we can project any video or graphics we wish onto the two projection screens without any concern of the effects on the silhouette extraction process. Also, standard fluorescent room lighting does not emit much infrared light and thus can be used to illuminate the environment without interfering with the infrared system. Our current system uses 5 inexpensive fluorescent spot lights, which are attached to the top of the screens.

This silhouetting process is illustrated in Fig. 2. Figure 2(a) shows a standard camera view of someone standing in front of the back-wall projection screen with graphics displayed. In Fig. 2(b), we see the same scene from the camera but now with infrared light being shown from behind the screen using the 6 infrared light 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 Fig. 2(c). The IR light is not visible to the human visual system and thus one sees only video projected on the display screen (as shown in Fig. 2(a)).

This method overcomes the color dependencies associated with chroma-keying approaches because it is based on the blocking (or eclipsing) of specialized light (see Fig. 3) rather than the color differences between the person and background. In our system, the person is always able to wear arbitrarily colored clothing. Furthermore, chroma-key and background subtraction systems require careful

[1] Many video cameras have an infrared-block filter which limits the use of this process. One may need to remove this filter, or use a camera without this filter installed. We used a Sony SSC-M370 black and white camera which passes much of the IR light.

[2] If the camera were placed above a screen or on the floor, the silhouette would be a bit more distorted from perspective effects. Also, the rear screen through which the infrared light passes is only slightly diffusive, and thus an off-center video camera would not register fully the light coming from the multiple infrared emitters spaced behind the screen.

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

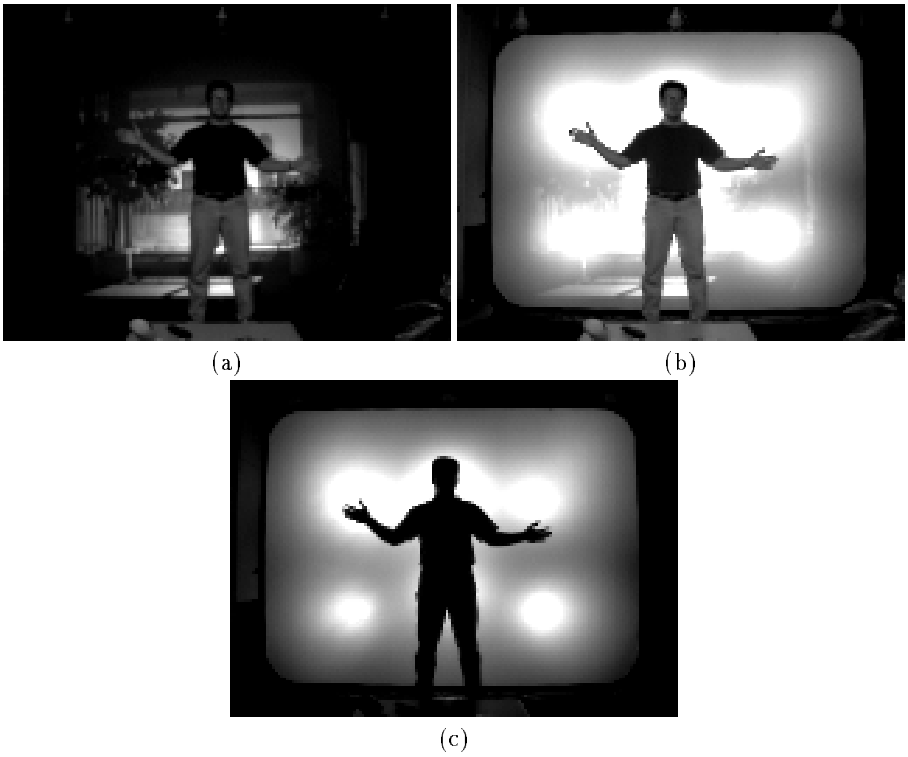


Fig. 2. Infrared light. (a) An image of the person in front of a video projection screen in the environment. (b) The same as shown in (a), but with infrared light directed through the screen. (c) The same image, but now filtered through an infrared-pass/visible-block filter. The image in (c) no longer shows the video projected on the screen, and the person now appears as a silhouette. To the naked eye, the version shown in (b) would appear as (a).

control of environment lighting, whereas the IR system is insensitive to arbitrary visible light. In comparison to systems that use bright rear-lighting, this system is similar but *hides* the technology from the participant by using the non-visible part of the light spectrum. The method also permits the display of video graphics behind the user, unlike the rear-lighting systems. Because the subject is rear-lit with the camera in front, any the reflection or absorption of the IR light occurs toward the rear screen, away from the camera. Therefore any hair, clothing, and material on the person that may cause reflective problems do not influence the imaging system.

We note that this system could be employed to the meteorologist scenario in the TV studios. Currently, a blue-screen (or green-screen) method is used to extract the meteorologist and place his/her image into a high-resolution weath-

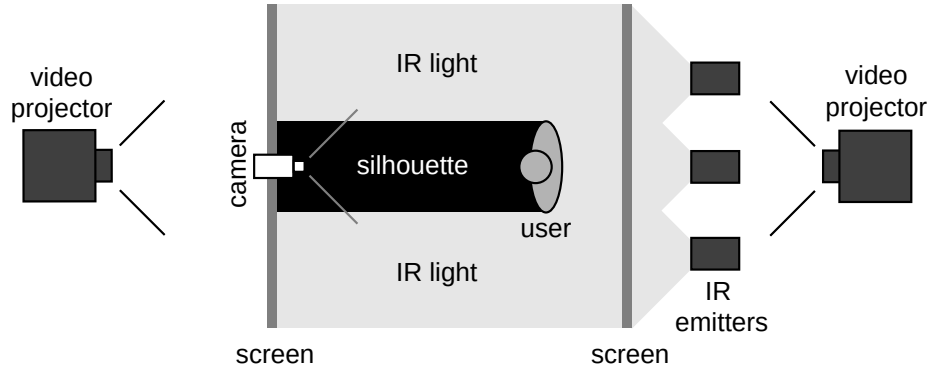


Fig. 3. Conceptual drawing of blocking (eclipsing) infrared light from the camera to generate a silhouette of the person.

ermap. The meteorologist must look off-camera to a remote monitor to view the resultant composite and see if his/her hand is over the correct region. With our approach, it is possible to accomplish the same composite result, but now have the added benefit of projecting the actual weathermap onto the back wall to help the meteorologist.

It is also possible to have another camera and side-screen with its own infrared emitters to recover an additional silhouette of the user as viewed from the side (see Fig. 4). Additional information (e.g. three-dimensional information) of the person could be attained using the two silhouettes (one from the front and one from the side). The side-screen infrared camera/emitters would need to be tuned at a different wavelength, modulation, or synchronization than the back screen camera/emitters as not to interfere with each other.

2.3 Image Processing

The advantage of creating a robust silhouette image of the person using the above lighting approach is that we can use simple image processing methods to easily and quickly (in real-time) extract the silhouette from the digitized video. We could use a simple thresholding of the image to find the silhouette, but the emitters are not widely diffused by the projection screen and there are varying degrees of brightness (i.e. the IR light is not uniformly distributed across the screen as shown in Fig. 5(a)). Instead, we chose to follow the common background subtraction methodology [3, 15, 7], where first a reference picture is taken without a person in front of the screen (see Fig. 5(a)). Then for any new image containing the person (see Fig. 5(b)), all the pixels in the screen area are compared between the reference image and this new image, where a pixel is marked as belonging to the person if the difference between the reference and current image at that pixel is above some threshold (see Fig. 5(c)). Due to

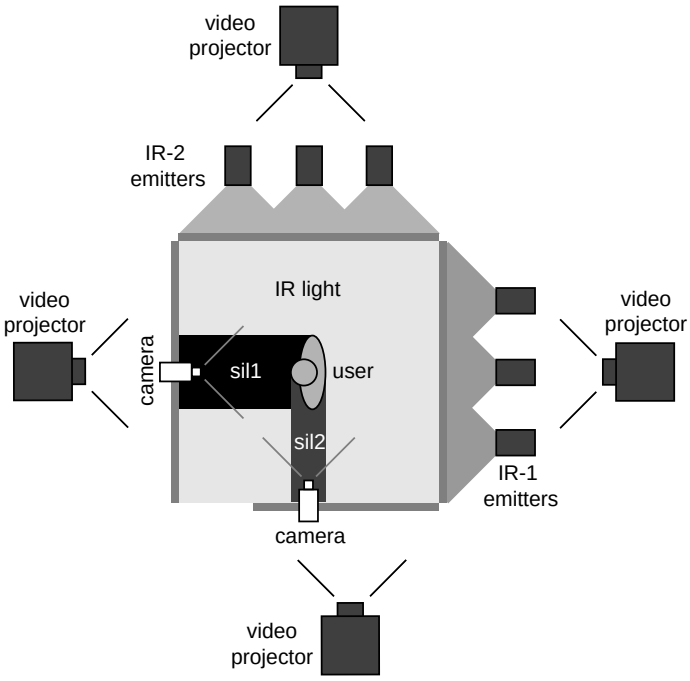


Fig. 4. Multiple screen environment using two independent IR emissions.

imaging noise and digitizing resolution (in our case, 160x120), spurious pixels may be set and small thin body regions may end up disconnected from the rest of the body. We can apply simple image morphology (dilation) to the difference image to re-connect any small regions which may have become disjoint, and then perform simple region growing to find the largest region(s) in the image [6] (see Fig. 5(d)). This retains only the person while removing the noise. The result is a slightly fuller silhouette of the person, which can be further examined using computer vision algorithms for measuring the location, posture, and motion of the person to drive the interaction.

3 Virtual Aerobics

In this section we discuss the design and implementation of a *virtual Personal Aerobics Trainer (PAT)* employing the above IR silhouetting environment[4]. The aerobics application demonstrates the usefulness and capability of the IR silhouetting approach.

[4] An extended description of the virtual aerobics system can be found in [5].

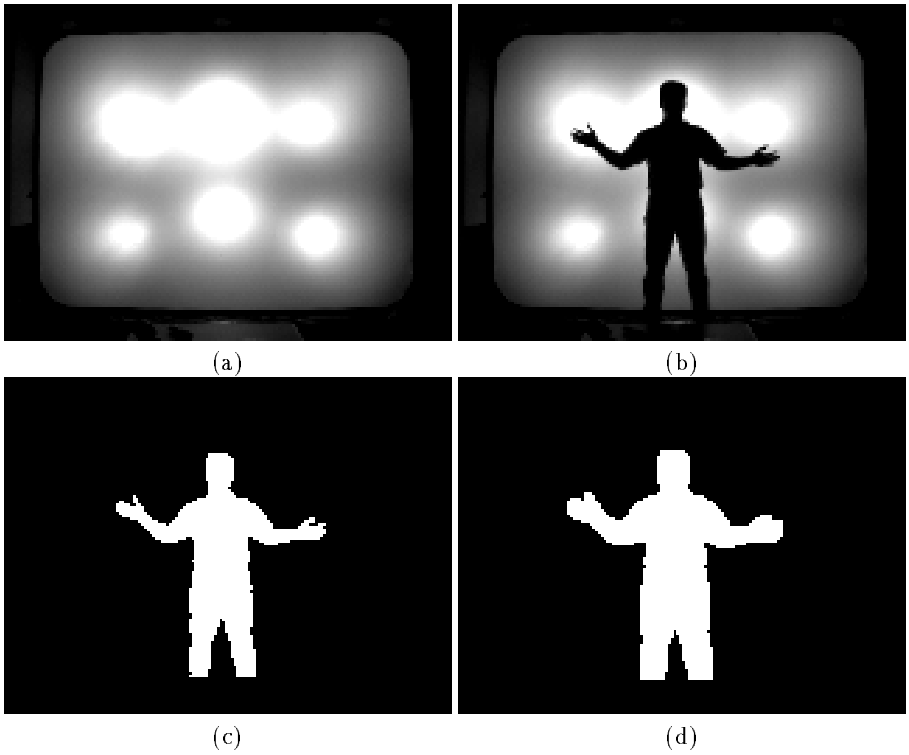


Fig. 5. Image processing. (a) Reference image. (b) Input image. (c) Binarized difference image. (d) Image morphology and region growing result.

The PAT system creates a personalized aerobics session for the user and displays the resulting instruction on a large front screen (or TV monitor). Here the user can choose which moves (and for how long), which music, and which instructor are desired for the workout. The session created by the user is then automatically generated and begins when the user enters the area in front of the screen (see Fig. 6).

The user periodically receives audio-visual feedback from the virtual instructor on his/her performance. To accomplish this, we use the silhouette form extracted by the IR system and use real-time computer vision techniques to recognize the aerobic movements of the user from the silhouette. Based upon 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 correctly, or "follow me!" if the user is not performing the move correctly). When performing large-scale aerobic exercise movements, having a wireless interface (e.g. no wired body-suit) enables the experience to be more natural and desirable [2, 15, 12].

Fig. 6. Virtual PAT. A virtual personal aerobics trainer. Photo credit: Webb Chappell. Copyright: Webb Chappell 1998.

The underlying motivation for building the virtual aerobics 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 room just knew whether or not a person was there moving in front of the screen. 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).

3.1 System Design

The PAT system is a modular design of media and vision components. All software[5] was written in C++ and run on SGI R10000 O2 computer systems (though we believe all the components could be placed within a much lower-end hardware setup). The output video of the system is sent to the frontal screen, as shown in Fig. 6, showing the instructor performing the moves. The feedback is

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

Fig. 7. Video output of virtual instructor (an Army Drill Sergeant).

routed through the audio channel. The music is currently in the form of MIDI files.

Currently, a set of movie clips, showing a full view of the instructor, is used (see Fig. 7). 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[6].

Each time a new aerobic move begins, 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[7], it randomly picks and outputs 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

[6] The speed of the current movie clip can be altered to be in synchronization with the MIDI music currently being played.

[7] The system checks every movement cycle to see if the user is complying. A *negative* comment is given immediately (every cycle) until the user performs the move correctly. If the user is performing the move, a *positive* comment is given at predetermined intervals (e.g. every few cycles or every few seconds).

for the system, which increases the entertainment value of the system as well as its usefulness.

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 (e.g. the lips of the instructor do not move, as if speaking the lines). One could consider 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 be 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 and audio clips for simplicity.

3.2 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 [12]). 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. While we currently use only one instructor (a brash Army Drill Sergeant), the system is designed to have multiple instructors from which to choose. The program is instantly available upon instantiation of the system with a script, and is not generated off-line (not compiled). The system loads the script and initiates the program when the user enters the space. Currently the script is modified in a text editor, but its simple form would make the construction of a GUI script selector trivial.

3.3 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. 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 pause or 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.

3.4 Recognizing Aerobic Movements With Computer Vision

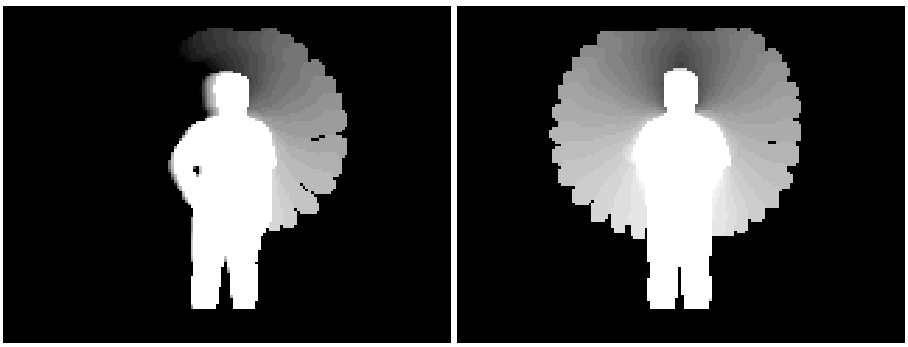
Real-time computer vision techniques developed by the authors were used to "watch" the user and determine if he/she is performing the same move as the instructor.

The first task of the vision system is to monitor the area and make sure someone is actually present in the space. This is easily accomplished by looking for the presence of the person's silhouette generated by the IR system. The PAT system then starts-up when a person enters the space. Also, if the person prematurely leaves the area during the workout session, the system recognizes that the person has left and correspondingly shuts-down or pauses the session.

Recently, we have developed computer vision methods which show promising results in recognizing such large-scale aerobic exercise movements [4]. 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 a motion template, Fig. 8 shows the templates generated from the IR silhouettes for the movements of left-arm-raise (left-side stretch) and fan-up-both-arms (deep-breathing exercise stretch). Training data of each of the moves executed by several users are collected to get a measure of variation which may be seen across different people. Statistical pattern recognition techniques are then employed for the recognition task. This approach easily extends to multiple camera views of the person. To ease in discussion here, we point the reader to [4] for details on the algorithm.

4 Summary

In this paper we presented a simple and robust method for extracting a silhouette of a participant, overcoming the inherent problems associated with using traditional chroma-keying, background subtraction, and rear-light projection methods. The resulting system also makes available a new venue for multi-screen interaction by incorporating multiple video screens without requiring any special clothing or wired technology. We showed how a robust silhouette of the user can be extracted using specialized infrared lighting without making the underlying technology apparent to those interacting within the environment. To show an example application using the system, a virtual aerobics instructor was presented. The aerobics system applies special computer vision methods to the infrared silhouette to recognize the movements of the user. This then guides its interaction with the participant. The infrared sensing framework itself is a portable system which can act as a re-usable infrastructure for many interactive projects.



(a) (b)

Fig. 8. Example motion templates for IR silhouettes. (a) Motion template for left-arm-raise (left-side stretch). (b) Motion template for fan-up-both-arms (deep-breathing exercise stretch).

5 Acknowledgments

We first would like to acknowledge discussions with Joe Paradiso and Thad Starner for their helpful advice regarding the use of infrared light. We also would like to thank Andy Lippman for playing out the role of the Army Drill Sergeant in the virtual aerobics application. We lastly acknowledge the support of the Digital Life Consortium at the MIT Media Laboratory for this project.

References

1. Blumberg, B., "Old tricks, new dogs: ethology and interactive creatures," PhD dissertation, MIT Media Lab (1996)
2. Bobick, A., Intille, S., Davis, J., Baird, F., Campbell, L., Ivanov, Y., Pinhanez, C., Schutte, A., Wilson, A., "The KidsRoom: action recognition in an interactive story environment," *Presence* (to appear)
3. Darrell, T., Maes, P., Blumberg, B., Pentland, A., "A novel environment for situated vision and behavior," *IEEE Workshop for Visual Behaviors* (1994)
4. Davis, J., Bobick, A., "The representation and recognition of human movement using temporal templates," *Comp. Vis. and Pattern Rec.* (1997) 928–934
5. Davis, J., Bobick, A., "Virtual PAT: a virtual personal aerobics trainer," *Workshop on Perceptual User Interfaces* (1998)
6. Gonzalez, R., Woods, E., *Digital image processing,* Addison-Wesley (1992)
7. Hogg, D., "Model-based vision: a paradigm to see a walking person," *Image and Vision Computing*, **1** (1983)
8. Ishii, H., Ullmer, B., "Tangible bits: towards seamless interfaces between people, bits and atoms," *Conference on Human Factors in Computing Systems* (1997) 234–241
9. Krueger, M., *Artificial reality II,* Addison-Wesley (1991)

10. Murray, J., *Hamlet on the holodeck,* The Free Press (1997)
11. Paradiso, J., "Electronic music interfaces," *IEEE Spectrum* **34** (1997) 18-30
12. Pinhanez, C., Mase, K., Bobick A., "Interval scripts: a design paradigm for story-based interactive systems," *Conference on Human Factors in Computing Systems* (1997) 287–294
13. Rekimoto, J., Matsushita, N., "Perceptual surfaces: towards a human and object sensitive interactive display," *Workshop on Perceptual User Interfaces* (1997) 30–32
14. Strickon, J., Paradiso, J., "Tracking hands above large interactive surfaces with a low-cost scanning laser rangefinder," *Conference on Human Factors in Computing Systems*, (1998) 231–232
15. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A., "Pfinder: real-time tracking of the human body," *SPIE Conference on Integration Issues in Large Commercial Media Delivery Systems,* (1995)