M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 457
Vision and Modeling group

SIDEshow: A Silhouette-based Interactive Dual-screen Environment

James W. Davis **Aaron F. Bobick**
MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139
(jdavis | bobick@media.mit.edu)

Abstract

In this paper, we present a new multi-screen interactive environment. The system extracts a silhouette of the participant for driving the interaction using a method that overcomes the inherent problems associated with traditional chroma-keying, background subtraction, and rear-light projection methods. We present an approach for generating a robust silhouette of the participant using specialized infrared lighting while not making the underlying technology apparent to those interacting within the system. The design also enables video projection screens to be placed in front of and behind the user without interfering with the silhouette extraction process. The framework itself 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 [9, 7, 8, 1, 2]. As for sensing the person in the space, some installations use specialized light, lasers, or electric field sensing to detects hands or objects [9, 7, 8, 4]. Other approaches use similar variants of chroma-keying (i.e. blue-screening) [1], background subtraction [10, 3], or rear-light projection [5] to enable a video camera to extract a silhouette the person, where the person need not wear special clothing. We are interested in the latter approaches where a full-body silhouette is extracted and then used as input for driving the interaction, where the position, shape, or motion of the silhouette is of importance.

The main problem with the current technology for extracting the silhouette is that it relies almost solely on the color components of the video signal to perform the extraction. For example, chroma-keying methods have the person stand in front of a background consisting of a uniform-colored wall or screen. Then, 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 by detecting and removing all the background color in the image (leaving only the person in the image). There now must be a restriction placed on the user to not have the color of the background anywhere on his/her clothing. If the space is to be used as an interactive environment, the color methods as well as the rear-light approach have an obvious technological crutch on display for all to see. Immersion is a strong requirement when building interactive environments [6], and part of this illusion may be broken if such walls are incorporated.

One slight variant of the chroma-keying method is known as background subtraction. Here a snapshot of the environment with no people inside is stored as a reference image. Then, any video image seen later (containing perhaps a person) is compared with the reference image. Anything that differs, at a pixel-wise level, is assumed to belong to an object/person in the environment. A more natural scene, rather than just a colored or rear-light wall, can be used. For this to work, the colors on the person still need to be *different* than the wall/objects behind them. As long as this constraint is held, 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), there are problems that arise distinguishing the background from the person. 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 (Section 2). Next, we discuss possible applications which could use such an environment (Section 3), and then conclude with a brief summary of the framework (Section 4).

2 Design specification

We divide the specification of the system into three main components. First, the environment itself is examined as an interactive space. We next present how specialized lighting can be used to enable robust sensing of the participant. Lastly, simple computer vision techniques are shown for extracting the silhouette from the video stream.

2.1 The environment

The layout of the space contains two large video projection screens, with one behind and one in front of the user (See Figure 1), with the primary screen as the frontal video display. Here, video or graphics can be displayed on the two screens, enabling virtual objects or interactions on either, or both, of these displays[1]. 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). These collapsible screens also allow the system to be easily portable to different locations, unlike the transport of large projections 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 dictates this pleasing viewing distance from the front screen.

In our design we imposed a simple room theme and placed a coffee-table, some plants, and a chair inside the space to give a more natural feel of being in a room. By restricting the user to move within a zone around the back screen, we can place objects in the front part of the environment which can then be hidden from the camera's view of the person (shown later). By furnishing the environment with objects that fit the application (in our case, we wished to recreate a household room), we remove the sterility found in those environments composed strictly of large screens.

(a)

(b)

Figure 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.

2.2 Infrared lighting

As for extracting a reliable silhouette of the user with a live video screen behind, we begin the process by directing invisible infrared light (using 6 consumer 840nm emitters) through the large back-projected screen behind the user (See Figure 2). By using an infrared-pass/visible-block filter tuned for this wavelength of infrared light, we can then restrict a black-and-white video camera[2] placed in front of the user to see only this infrared light. Since a person standing in front of the screen physically blocks the infrared light coming out of the screen, a video camera placed in front of the user therefore sees a bright image (of the infrared light passing through the screen) with a black silhouette of the person (from where the person is blocking the infrared light to the camera). To get the most flat, frontal view of the person, the camera needs to reside approximately hip-level to the user[3]. Because the camera can't sit behind the front screen (with the screen blocking the view, and the camera causing shadows 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 in the space, and still provides an adequate virtual window.

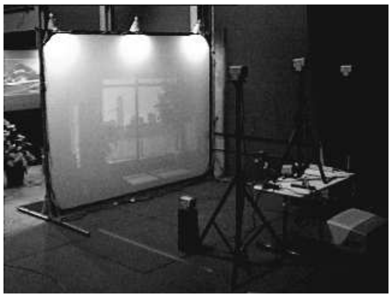


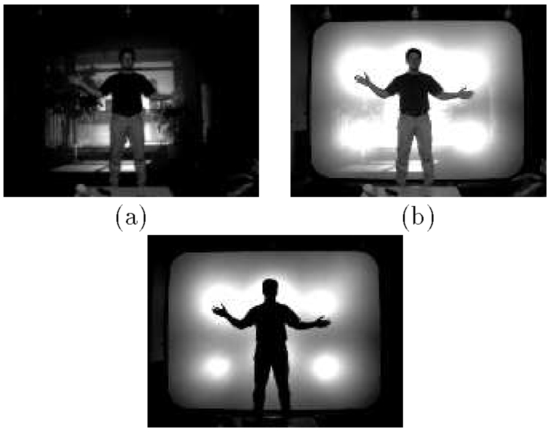
Figure 2: View from behind the environment showing the infrared emitters aimed at the rear screen.

Since the system is tuned for infrared light and the fact that many video projectors emit very little infrared light (or can be outfitted to do so with infrared-block filters), 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, fluorescent room lighting does not emit much infrared light and thus can be used to illuminate the environment without interfering with the infrared system. It is conceivable to have another camera and

[1] No real side wall (as shown in Figure 1) is necessary. It is possible with the approach to place another video screen to act as the side wall for additional virtual interactions.

[2] Many video cameras have an infrared-block filter which limits the use of this process. One should use a camera without this filter installed.

[3] 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.



(a) (b)

(c)

Figure 3: 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).

side-screen with its own infrared emitters[4] to get a silhouette of the user as viewed from the side. This way some three-dimensional information of the person can be attained using the two silhouettes.

We can see this silhouetting process in Figure 3. Figure 3(a) shows a camera view of someone standing in front of the back-wall projection screen with graphics displayed. In Figure 3(b), we see the same picture but now with infrared light being shown from behind the screen using 6 infrared light emitters[5]. By placing an infrared-pass/visible-block filter over the video camera lens, a brightly lit screen (with no graphics visible) and a silhouette of the user is seen.

This method overcomes the color dependencies associated with chroma-keying approaches because it is based on the blocking (or eclipsing) of specialized light (See Figure 4) rather than the color differences between the person and background. In our system, the person is able wear any color clothing, and the end effect is the same (unlike the color-based methods). In comparison to systems that use bright rear-lighting, this system is similar but hides the technology from the participant by using a different part of the light spectrum. The method also permits the display of video graphics behind the user, unlike the rear-lighting systems. As with rear-lighting, because the camera is in front of the person with the light coming from behind the person, any odd reflections that may occur on the person from the infrared light point away from the camera, and therefore any clothing or material on the person that may cause reflective problems does not influence the imaging system.

[4] The side-screen infrared emitters need to be tuned at a different wavelength or modulation than the back screen emitters as not to interfere with the backlighting process.

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

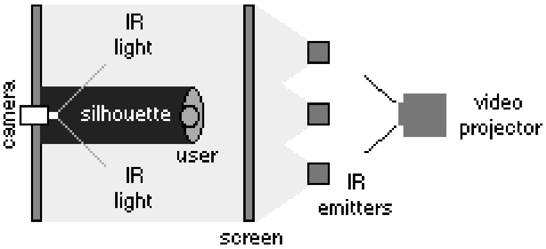


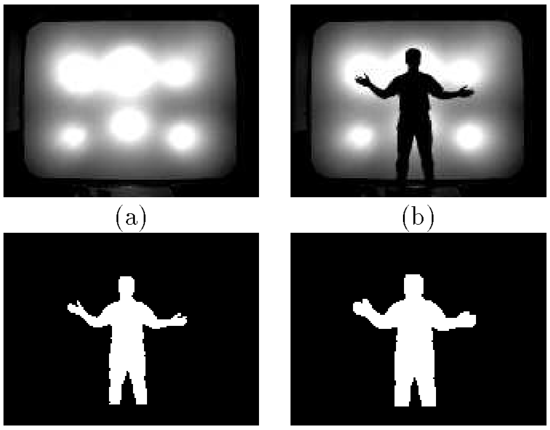
Figure 4: Conceptual drawing of blocking (eclipsing) infrared light from the camera to generate a silhouette of the person.

2.3 Computer vision

Since the effort has been made to create a robust silhouette image of the person using the above lighting approach, we can use simple computer vision methods to easily and quickly extract the form from the digitized video. To begin 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. Instead, we chose to follow the common background subtraction methodology, where first a picture is taken without a person in front of the screen (See Figure 5(a)). Then for any new image (See Figure 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 Figure 5(c)). Due to noise and the digitizing resolution (160x120), spurious pixels may be set and small thin body regions may be disconnected from the rest of the body. We can apply simple image morphology (dilation) to the difference image to connect-up any small regions which may have become disjoint, and then perform simple region growing to find the largest region(s) in the image (See Figure 5(d)), thus retaining the person while removing the noise. The result is a silhouette of the person extracted in real-time, which can be further examined using computer vision algorithms for measuring the location, posture, and motion of the person.

3 Applications

Because there is a screen behind the participant in the space, we can also engage the passive users outside of the environment (perhaps waiting their turn to go inside). Though the participant inside may have direct control over what happens on the forward screen, additional graphics can be displayed on the back screen to more fully immerse the participant and also to help pull in and engage the people watching from outside of the space.



(a) (b) (c) (d)

Figure 5: Computer Vision. (a) Reference image. (b) Input image. (c) Binarized difference image. (d) Image morphology and region growing result.

The framework we have presented is general enough for the system to used in a variety of applications. In Krueger's well known work [5], he mainly relied on a silhouette of the participant to drive the interaction. A phosphorescent light-wall behind the user was employed to extract a clean silhouette. Our system could be easily ported to his environments, where the extraction process works just as well but is hidden from the participant and offers an additional interaction screen.

We have recently begun to develop our own interactive project in this space. To incorporate our additional research efforts of recognizing human movement, we developed a prototype system for a virtual Personal Aerobics Trainer (virtual PAT) [2]. 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 virtual instructor (See Figure 6) by enabling it to watch and talk to the user (instead of just the user watching the screen). Presently, the sensing for the system is based on the silhouette of the person, as generated by the framework discussed in this paper.

In general, such an environment as described is well-suited for a variety of single or multi-screen interactive projects. The additional screen behind the user, while necessary, can also be used to enhance and make the experience more rich and immersive to the participants.

4 Summary

In this paper we presented a method to overcome the inherent problems associated with extracting a silhouette of a participant 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 video screens in front of and behind the participant. 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 system. The framework itself is a portable system which can act as a re-usable infrastructure for many interactive projects.

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

5 Acknowledgments

We first would like acknowledge discussions with Joe Paradiso and Thad Starner for their helpful advice regarding the use of infrared light. We also acknowledge the support of the Digital Life Consortium at the MIT Media Laboratory for this project.

References

[1] Darrell, T., P. Maes, B. Blumberg, and A. Pentland. A novel environment for situated vision and behavior. In *IEEE Wkshp. for Visual Behaviors (CVPR-94)*, 1994.

[2] Davis, J. and A. Bobick. Virtual pat: a virtual personal aerobics trainer. MIT Media Lab Perceptual Computing Group Technical Report No. 436, MIT, 1997.

[3] Hogg, D. Model-based vision: a paradigm to see a walking person. *Image and Vision Computing*, 1(1), 1983.

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

[5] Krueger, M. *Artificial reality II.* Addison-Wesley, 1991.

[6] Murray, J. *Hamlet on the Holodeck: the Future of Narrative in Cyberspace.* The Free Press, Simon & Schuster, 1997.

[7] Paradiso, J. Electronic music interfaces. *IEEE Spectrum*, 34(12):18–30, December 1997.

[8] 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.

[9] Strickon, J. and J. Paradiso. Tracking hands above large interactive surfaces with a low-cost scanning laser rangefinder. In *Submitted to CHI'98*, 1998.

[10] 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.