Vision-Based Multiple-User Interaction with In-home Large Displays

by

Wei You

B.Eng., Nanjing University of Science and Technology, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

The Faculty of Graduate Studies

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August, 2008

Abstract

In-home large displays such as TVs are becoming larger in size, and more interactive in function. They start to be simultaneously used by multiple people for various tasks in a dynamic setting. User interface issues such as multiple users sharing the screen resources of the displays, and the usage of multiple control devices have begun to emerge. We assume horizontally laid out "personal interaction spaces" as the user interface for multiple users to manage their screen real-estate. In this case, users often need to sign in and out as well as have their personal spaces placed on the screen. Also, the limited number and complex usage of conventional remote controllers for TVs cannot satisfy the need of multiple interacting users.

In this thesis, we consider a computer vision based system as a solution to the emerging user interface issues. We built a vision system that tracks the identities, positions and hand positions of people in front of a large display to support our user studies of screen real-estate management and multi-device management. We explore the usefulness of a vision system through two user studies.

We designed the first study to compare the use of tracker-based mechanisms versus manual ones for managing the display. Study Results suggest that the tracking system is especially useful for simplifying the user sign-

in/out process in conjunction with a manual method, and effective user-centric placement of people's interaction spaces.

As well, we designed a second study to explore whether contexts exist for lower fidelity, gesture-based "remote controllers" for manipulating on-screen objects. Study results show that gestural interfaces combined with high fidelity devices such as a mobile phone in a group gaming scenario can be useful for centralizing the control in the team and reducing errors. However, gestural control is only suitable for simple, once-in-awhile interaction.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors Dr. Sidney Fels and Dr. Rodger Lea for their guidance and continual help from my academic research to my working habit over the past two years. Thanks to Dr. Fels for all the ideas we have discussed, and all your patience with me in every detail in this work. Thanks to Dr. Lea for your meticulous and timely advice on my work.

Thanks to Panasonic R&D team who have been supporting my research, and I really appreciate all the inspiring discussion we have had.

I have learned a great deal from and am very grateful to my fellow researchers and students from both Human Communication Lab (HCT) and Media and Graphics Interdisciplinary Center (MAGIC) who selflessly provided me with help and advice throughout the project. Thanks to Dr. Hao Jiang who helped me at early stage of the project and inspired my interest in computer vision; to Dr. Mattias Finke, Mike Blackstock and Changsong Shen for offering great advice when I got stuck; to Meghan Deutscher and Donovan Parks for going through the user study with me with patience; to Tony Tang as a great reader of my thesis; to Nicole Arksey and Nels Anderson for your encouragement in my first year. I am also very grateful to all others in the lab for indulging me with endless request for tests and user studies.

Thanks to my fellow residents in St.John's College. My life in Vancouver would not have been as happy without all of you as company. Special thanks to Colin Doutre and Michelle Tan for helping to proofread my thesis, and a lot others for taking part in my user study.

And to my dearest family, who have always being loving me and supporting me no matter what happens. Thanks to my dad Li You, who have never failed to encourage me and have passed on his enthusiasm in engineering to me. To my mom Beiwei Feng, who is so smart and caring in guiding me through all the difficulties in life. Thanks to my wise and loving grandpa Shichang Feng and grandma Jie Lu - growing up with you in the university has inspired my endless pursuit for knowledge and education. Thanks to my cousin Chao Wang, who has been taking care of me ever since I parachuted in Vancouver; and to Jingyuan Li, who has always been standing by my side over the past six years. Without the love and support of all of you, life would have been much more boring, and this work would have been much more difficult.

There won't be enough thanks to those in my life who care for me and believe in me. I hope you can all see my growth.

Chapter 1

Introduction

Homes are evolving to accommodate smart interactive environments consisting of a networked collection of electronic devices and displays. Compared to all other displays at home, TV screens are larger in size, and are typically located in the central areas such as the living room where family members and friends socialize. Therefore they are very important for entertainment activities at home. Manufacturers are trying to build TVs of larger sizes in order to maximize the entertainment experience, and the families are happy to purchase these large screen TVs. For example, our collaborator Panasonic has made a 150" TV which is the world's largest flat panel TV [3]. A survey done by Quixel Research [34] regarding TV usage finds that consumers have the space, opportunity and significant budget to purchase a large screen TV: 77% of consumers surveyed would like to have a screen size larger than 40 inches.

At the same time, TVs are being developed to be more versatile in function than to simply display TV programs or videos, but also contents that allow and encourage user input from various types of electronic devices, such as games. Thus, the activities that users will be involved in with the TV are no longer confined to passive TV watching, but may in-

clude interactive activities, such as organizing media files, browsing the web, sending and receiving messages, gaming, and so on. Panasonic has built an in-home wall-size interactive display Life Wall [3] that supports watching video, browsing Google Earth, and doing video chat with friends. Further, a TV of increased size allows multiple people to perform complex interaction with it concurrently. They can either use part of the screen for their own activity, or engage in group activities, such as playing video games as a team or manipulate family digital media contents (such as photo albums or notes) together at the same time. The increased size and enhanced capabilities of TV screens suggest that in the near future, in-home large screen TVs may start to become the center of a wide range of interactive activities in home entertainment, which means multiple users will be likely to share the display for different tasks at the same time.

Due to the emerging interactive nature of large display TVs, we consider the context of this thesis to be a class of futuristic displays: a combination of large, high resolution TVs and computer displays. This class of displays have remote controllers, and are attached to processors, which allow multiple users to perform interactive activities on it. Throughout this thesis we use the terms "large in-home interactive display" and "large display" to mean the same thing.

Despite the changing trend of TV display usage, the predominant model of interaction for home consumer electronics is still based on remote controls instead of a smart environment augmented by sensors such as cameras and movement sensors. Most current interfaces with large display TVs focus on single user, single device interaction using a complex remote control.

While people at work often use a single large display for a common task, users at home are more likely to use the large display at the same time for different tasks. In this case, they will necessarily compete for screen real-estate. Arksey et al. [5] studied this emerging trend and showed that when in-home large displays get sufficiently large, people are comfortable sharing the real-estate for different tasks. An additional difficulty is that usage configurations are not static as those at work: people may come and go, which means screen real-estate needs are dynamic. Thus, issues regarding how to allocate the screen real-estate among multiple users to facilitate natural, intuitive use become a relevant, complex and interesting topic.

Complicating matters is that when multiple users share the use of a large display, conventional control devices are usually limited in number and complicated to use. We are in need of an interface that provides simple, cost-effective, natural, and ad-hoc remote control over the on-screen contents.

Smart homes technology may be able to help manage this complex new interactive environment as they can provide information about who is where and what they are doing in a given context. This information may be able to help decide whether a person wants to use some screen real-estate for his personal activities, and where that real-estate should be placed on the large, shared display. Further, smart environments may also be able to help with supporting different types of control techniques such as shifting and collaboration of input control from a household remote control to a user's hand gesture. As a starting point, we believe that mechanisms are required for a large display system to have information about who is doing what for persistence of activities, managing screen real-estate and helping

to manipulate on-screen contents.

Our collaborator Panasonic envisions that a computer vision system can provide such mechanisms to extract information about identities, locations and actions of people in front of the display. Based on this information, the smart display system can react according to the users' presence and potential need expressed by their actions. In addition, a vision system enables free, untethered interaction with the large display, and requires no more than additional cameras and software that processes the camera feed. Since it is unlikely that users will install a camera system separate from the large display, having cameras embedded in and coplanar to the large display is a more realistic setup from the manufacturer's point of view. Panasonic has confirmed this position as well. In this thesis, we will base our work and discussion upon this setup.

Figure 1.1 shows the mock-up of a vision-augmented display system. In the future, cameras can be manufactured in the large display, and users sitting on the couch can not only use various types of remote devices, but their own presence, and possibly their body postures and hand gestures to communicate their intention to the large display so as to control their own contents.

Computer vision based systems that help interpret people's identities, locations and actions have been extensively studied. However, how to build a system that supports the needs in a domestic large display environment, whether a vision system is sufficient for such home use, and what it can be used for remains unsolved. This thesis tries to address the above questions by building a vision based system for a large display, and carrying out user

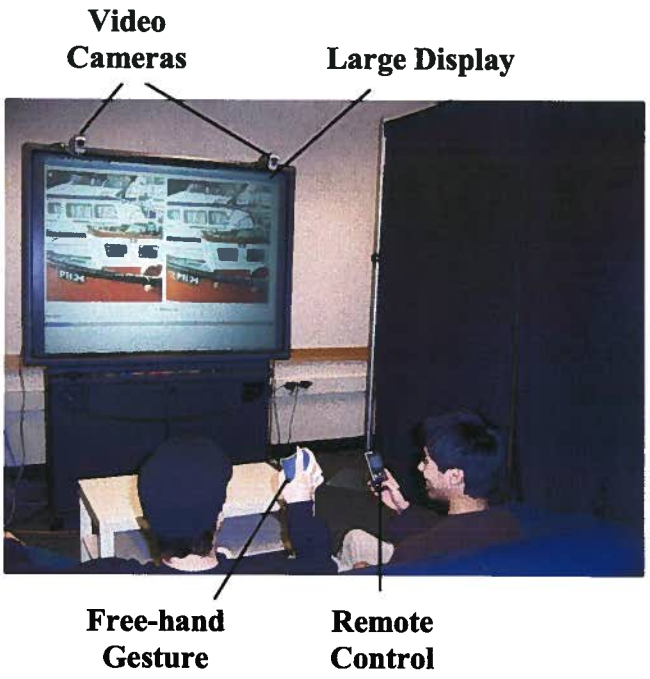

Figure 1.1: Multi-user Multi-device Interaction with a Large Display.

studies to investigate the usefulness of the vision system.

The investigated issues in the thesis is shown in Figure 1.2. It consists of two parts: the implementation and discussion of the supporting vision system, and the in-home large display user interface study based on the vision system. We built the vision system to address the user interface issues that arise from the scenario of multiple users interacting with in-home large displays that have cameras attached. The user interface study based on the vision system reveals the usefulness and lessons for building an applied vision system. The two parts converge in whether and what a vision system is useful for in the domestic display environment. Challenges and motivation for both parts will be discussed in Section 1.1. Section 1.2

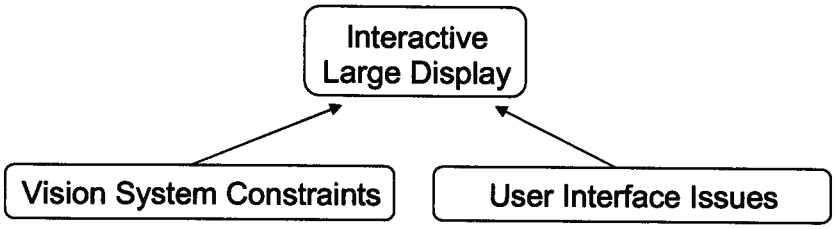

Figure 1.2: Investigated Issues in the Thesis. We study the challenges of the vision system and the issues of user interface in order to understand of the usefulness of a vision system in the interactive large display system.

describes our research approach including the basic user interface design. We give the research goals and hypotheses of this thesis in Section 1.3, and an overview of the thesis in Section 1.4.

1.1 Motivation and Challenges

In this section, we discuss the motivation of this thesis in two mutually dependent parts. The first part discusses the user interface issues needed to be addressed in the scenario of multiple users dynamically interacting with the large display. The second part identifies the challenges faced by the vision system given the constraints from interaction, the home environment, and product design.

1.1.1 User Scenarios and Interface Issues

The context of our discussion is an in-home large display that allows multiple users to perform interactive activities on it concurrently. We first present two imagined scenarios in this context, and then analyse some of the arising

user interface issues.

Scenario 1: Bob and Jane coming home.

Bob comes home, sits on the couch in the living room, and turns on the large plasma TV with the remote control. He switches the channel to his favourite hockey game and displays it in full screen. A few minutes later, Jane comes into the living room and sits on the couch, too. She is in a hurry to the local mall and wants to use the TV to display some shopping information from the Internet. However, Bob does not want to give up the program he is watching, so they use the remote control to divide the large screen into halves, one part displaying the hockey game and the other part the shopping information. After a while, Jane leaves for the mall. She does not care whether the shopping information is still there, but Bob wants the full screen to watch the game. So Bob expands his side of the screen to full screen using the remote control.

Scenario 2: Airplane Game.

Jane and Bob would like to play a picture matching game on the large screen TV together: spotting and marking the differences between a pair of similar pictures. They will spot the differences faster and have more fun if they play the game together. However, there is only one remote controller. So Jane and Bob need to pass the only controller back and forth if any of them wants to mark a difference he/she spots.

These two scenarios reveal several issues regarding the usage of a large display in a home-based multi-user simultaneous interaction context, listed as follows:

Whose contents to display

People are used to the notion of a television, which preserves the state of the display contents based on what was on the screen. Thus, when a person turns on the TV, it shows the most recently watched channel regardless of who was watching the channel. However, when multiple people share the use of the large display, a person does not necessarily see the contents he last watched when turning on the large display. It is even more troublesome if the contents displayed on the large display are not only TV programs, but also various other personal contents such as emails, or unfinished video games. For an interactive large display, we need an approach to manage multiple users' diverse need of information, so that when one person turns on the large display, the contents of his own interest are shown.

Displaying contents for multiple users simultaneously

As the display gets bigger, there is a larger chance that multiple users want to utilize the display space at the same time. When this happens, it is not clear how their contents are displayed. They have the options of displaying all users' contents on the common space of the screen so each individual have access to the whole screen space (Figure 1.3 Left), or dividing the screen into different areas where different users' contents are displayed (Figure 1.3 Right).

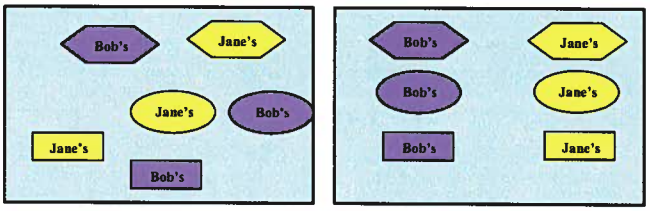

Figure 1.3: Left: Displaying all users' contents on a common space. Right: Displaying different users' contents in different areas.

Casual use in the home context

Unlike the workplace, where multiple people usually need to gather around a large display at the same time, people using a large display at home tend to behave in a more heterogeneous way both in time and in space: they come in and go out of the living room in different time, and have seats in different places of the room. They do not necessarily arrive to interact at the same time, and they do not need to leave at the same time either. In addition, while they are in front of the display, they could come in, go out and walk around, such as get up to grab snacks and drinks, or go to the washroom. Therefore, the casual use of the large display at home results in a much more dynamic multi-user interactive scene than the typically considered multi-user scenario at work.

More users than the number of physical controllers

When multiple users interact with the display simultaneously, the limited number of remote controllers might not be sufficient for the number of users. They need to either perform centralized control, namely having one person control the display contents, while everybody else sits there watching; or

token-passing floor control, which is to constantly pass the control among themselves to ensure everybody participates in the activity.

Complex interaction not suited for remote controllers

The interaction control necessitated by a large display may be complicated and requires frequent use. However, the keypad-based design of a conventional remote controller was primarily designed for functions in TV watching: switching the channels and changing the volume. When functions become more complicated, the combinations of key presses and possible mode changes may pose greater mental load on the users, forcing them to switch their attention from what they are doing. One extreme is to have an easy to use, multi-functional control device equivalent to mouse and keyboard to go with the large display, but the keyboard and mouse may not be suitable in an entertainment-oriented scenario. Novel controllers for video games, such as the Nintendo's Wiimote, Nunchuck, Wii Wheel [1] are emerging as promising control devices for interactive large displays, yet they normally come with the specific game console instead of being designed as universal control devices for the large displays.

Social aspects of sharing the displays

Performing collaborative tasks or playing collaborative games is an effective way of fostering social interaction among family members and friends. However, a conventional remote controller for everybody (if available) is able to perform all types of manipulation on the display, so users do not rely on each other in completing the task. Thus they seldom need to communicate when

taking part in group activities. It is promising to introduce controllers of different types so that users depend on each other's controllers to complete certain tasks or games, thus enhancing the social aspects of the interaction, such as participation, role-play and so on.

We consider two interface designs to address some of the issues regarding multi-user simultaneous usage of the large display, and focus our attention on two important issues: dynamically managing screen real-estate of the large display, and managing the use of control devices. The designs are described in Section 1.2.

1.1.2 Vision-based System Challenges

A vision-based system built for the smart large display interprets the scene in front of the display in real time, and provides information for the user interface to decide what contents to display. The major challenge of building the supporting vision-based system is to satisfy the needs of the smart display interface: to know the identities, positions and actions of the users in front of the large display. The challenges also lie in discovering whether and what a vision system can be useful for in this interaction context, which is a combined challenge with the user interface design.

The algorithm part that interprets the camera feed of the vision system should be able to identify registered people in front of the large display, and keep track of where they are, and what they are doing at the same time. This needs to be done in real time.

The algorithm that supports our user study needs to identify a small number of people (3 in our user study) and keep track of them and their

hands. The identities, body locations, and hand locations need to be fed into the interface application with an interactive rate for our user study.

Several constraints are imposed on the vision system from the viewpoint of product design and the home environment. The cameras are expected to be embedded and coplanar with the large display, so no visual information from other directions (such as from the ceiling) is provided. Further, this system needs to be purely vision-based; cues from other modals, such as speech, are not available. Also the cameras are supposed to be of consumer-electronics quality, without adding considerable cost to the display itself. In addition, the unconstrained lighting condition and dynamic user movements in a home requires the system to be robust to lighting changes, and to track multiple moving users accurately. For an interactive system, the vision component needs to work sufficiently accurately to be useful for enhancing user experience of interaction. Given all the challenges and constraints, accuracy of the system is of foremost importance to both a commercial system and the subsequent user interface study.

For the user study, some of the above constraints are relaxed. For example, the duration of a user study is much shorter than the duration of use in the home, which means the lighting does not change as much.

The Life Wall system [3] done by Panasonic is a wall-size display built with face and body recognition as an example of where a vision-based system can be used. There are other works that tried to build vision based systems to help the interaction experience with the displays. However, there is little work that discusses whether and how a vision system can help with user interfaces. We investigate this issue in the user study to inform future

design.

In summary, we are motivated to build a vision system attached to the display system that addresses the functional and non-functional requirements for the user study. From the user study we hope to generalize the usefulness of a vision system for in-home large display interaction.

1.2 Research Approach

In this thesis, we investigate two issues to understand the usefulness of a vision system in interaction with large in-home displays: managing screen real-estate, and managing control devices. Accordingly, we use two different user interface designs to solve the two issues respectively: a Personal Space Design for screen real-estate management, and a Complementary Device Design for input device management. We explain how we came up with these designs, and how they can lead to the discovery of usefulness of a vision system .

1.2.1 Personal Space Design for Screen Real-estate Management

This design is used to investigate whether people would like to have their personal media contents displayed directly in front of them on the large display. We assume that the large display is always in a neutral stand-by mode when nobody is interacting with it. We believe that it makes sense to group each user's personal contents since it is easy for him to find his contents on a display that is physically much larger than his own size. Therefore we

assign a "personal space" to each user with his identity (name or picture) on it, that contains his own contents, such as the last channel he watched, his personal photos, or emails, much like a window in the desktop environment.

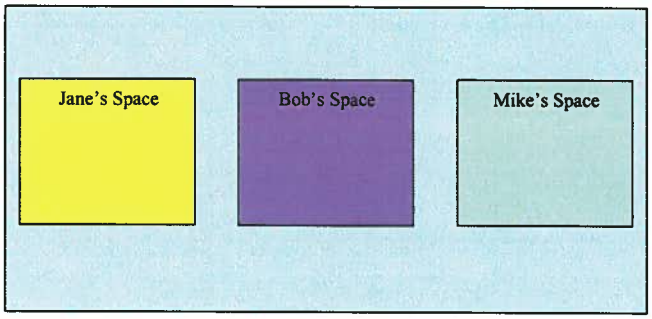

Figure 1.4: Personal Space Design.

When multiple users are interacting, we partition the screen horizontally, each partition containing one user's space. This conforms with the fact that users always stand or sit horizontally apart in front of the display. In this thesis, we always assume this fashion of sharing the screen real-estate among multiple users. An illustration of this design is shown in Figure 1.4.

Two further issues are reflected in this design:

1. **User registration.**

   For an interactive large display where everyone has his own personal space, even "turning on the TV" is a complicated process. In this design, "turning on the TV" means activating one's personal space that contains the state that you last left it in. To accommodate this requirement, a sign-in process is required when a user wants to "turn on" his own TV to allocate some of the screen to him.

Likewise, when a user leaves, and the space will no longer be in active use, a sign-out process is important to make room for others, and for privacy concerns. One obvious way for a user to sign in is to use a user-specific remote control, or a generic control that allows people to sign in with their names. We will explore whether an automatic vision system can make this sign-in and sign-out process simpler and more efficient than using a remote control.

2. **User space placement**

   When multiple people want to share the screen, some policies are required to establish where and how it is shared. A simple placement of user space based on, for example, when a person signs in, may cause disturbing sightlines that have to be corrected manually, disrupting the others. An automatic placement mechanism that senses where each interacting user is can help place the spaces according to relative locations of the users, so that the users do not need to cross sightlines in interaction. We try to find out whether a vision-based system is helpful in this respect.

We developed a vision-based system capable of identifying and tracking multiple users simultaneously. The output of such a system helps decide whose spaces should be shown, and where they should be. The vision-based system is described in Chapter 3. The first of the two user studies presented in Chapter 4 tests how multiple users experience managing their personal spaces with or without the help of the vision system. The results helps us learn the usability of this vision-based interface in managing large screen

real-estate.

1.2.2 Complementary Device Design for Device Management

This design is used to investigate whether gestural control can be a useful complement to conventional control devices in collaboration on the large display. Now feasible for in-home large displays, we consider the usefulness of gestures as a way that people can collaborate together. Technically, gestures are an attractive alternative since the vision system attached to the in-home system will be able to recognize them. Functionally, waving your hands and gesturing in front of the screen in a pre-defined way seems to be a natural method to manipulate on-screen contents.

However, gestural control is known to have the drawbacks of low recognition accuracy, tedious training, fatigue, unself-revealing, and so on. Therefore, we think it would better play an assisting role in group activities rather than acting alone. We hypothesized that we can separate some functions suitable for gesturing from the conventional remote controllers for gesture control to solve the problem of multiple people competing over the control devices in group activities and enhance cooperation at the same time.

We use a design in which multiple users collaborate in playing a game on the common display space with the two different configurations of devices: the Hybrid Control, and the Homogeneous Control. The Hybrid control configuration are devices with complementary functionalities: a remote controller, which is a high fidelity, master control, and low-fidelity free hand gesture control enabled by the vision system. The Homogeneous Control

configuration means each user has a remote controller. This design with the Hybrid Control configuration is shown in Figure 1.5.

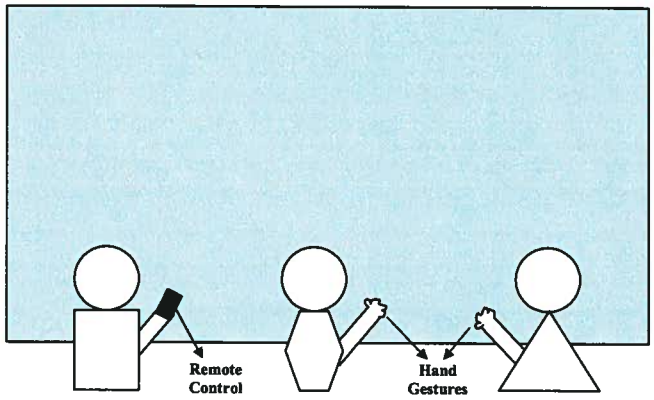

Figure 1.5: Complementary Device design (Hybrid Control Configuration).

The gesture control is built on top of the multi-user simultaneous tracking infrastructure. It allows users to wave a color pad to the camera to simulate free-hand gesturing. The second of the two user studies in Chapter 4 for device management tests how groups of users complete collaborative tasks with the two types of device configurations, in order to find out whether and how well gesture control can be useful as a complementary control device.

1.3 Research Goals and Contributions

The motivation and problems described in previous sections inspire us to build a vision system to address questions of whether home users would prefer to manage large screen real-estate manually versus automatically using a vision-based system, and how coarse-grain, low-fidelity control (i.e. hand gestures) can complement high-fidelity remote control devices, to simplify

control and enhance social aspects of interaction with the large display.

Based on our research approach, the goals of the thesis can be boiled down to a set of hypotheses:

1. Users prefer to be automatically signed-in and signed-out of his space rather than manually.

2. Aligning a user's workspace on the screen according to his relative horizontal physical location with other users helps improve his task performance and overall interaction experience.

3. Users cooperate better in the group of hybrid devices (a master control device and several low-fidelity gestural devices) than with homogeneous high- fidelity devices.

The contributions of this thesis are summarized as below:

- We built a proof-of-concept environment to investigate interactions in this emerging in-home space.

- We demonstrated the viability of people tracking and hand tracking in a relatively unconstrained environment.

- We found that a combination of vision-based automatic control and manual control can be useful for managing large display real-estate in multi-user interaction.

- We found that gestural control, with a high-fidelity remote controller, can help centralize control in group interaction with large displays, but can only be used occasionally.

1.4 Overview of the Thesis

We discuss the related work in Chapter 2. Chapter 3 describes the algorithmic structure of the multi-user identification and tracking system that is used in the interactive in-home environment in front of the large display. Chapter 4 presents the user studies we conducted, including the applications, study design, and report of the results and implications. We discuss the conclusions and future direction of the work in Chapter 5.

Chapter 2

Discussion of Related Work

Our thesis topic spans a number of areas including face recognition, real-time human tracking and hand detection in computer vision, communication architecture, mobile platform programming, gesture interfaces, and collaborative gaming on large displays. Because we focus on whether and how a vision-based infrastructure can be useful for multi-user interaction with in-home large displays, we first discuss related work on computer vision based smart systems. This part looks at systems that reacts according to interpreted location information of users, and systems that facilitate hand gesture interaction. Then we discuss systems and findings regarding collaborative work on a single display, which deal with issues of managing screen space and managing input devices. Finally, we examine work that looks at the usefulness of a vision system for interaction with smart environments.

2.1 Computer Vision Based Smart Systems

Vision technology has long been a candidate for enabling human computer interaction. For example, visual tracking may be used for analyzing large-scale (body, torso or arm) movements as well as small-scale (head, eye, finger) movements so that more natural, intuitive interactions can be made

available. We review some representative previous work that uses vision technology to provide interfaces for computer systems. Interaction based on location information of people will be discussed first, followed by a review of gesture and posture based interaction.

2.1.1 Location Information Based Interaction

Video tracking provides location information of users to serve as input to an interface and context information in a smart computing environment. In the case of interacting with a large display, full-body tracking information can be used to indicate presence, proximity, and relative positions to the display and with each other when multiple people are in the camera view.

Vogel et al. [39] demonstrate a public interactive display that reacts to identity, orientation, location and hand gesture information of users. Users are required to wear markers during the interaction, so that their motion information can be collected by a Vicon [4] motion tracking system. Multi-user interaction is supported by registration of markers. Their work proposes a framework for interaction phases: Ambient Display Phase, Implicit Interaction Phase, Subtle Interaction Phase and Personal Interaction Phase. In each phase, the contents on the large display and interaction modes change based on the location and orientation of users. For example, public information is displayed in Ambient Display Phase when nobody is interacting; notification is displayed in Implicit Interaction Phase when a user passes by; personal information can be navigated using hand gestures in Subtle Interaction Phase when users are close enough and pause for a moment; direct touch interaction is available for manipulating personal information in

Personal Interaction Phase when users are up-close. Users are able to make transitions between different phases by stepping closer or farther away. Personal information is displayed within users' body range, and users are also able to move their bodies or hands laterally to navigate information in the Subtle Interaction Phase.

Similar to Vogel et al.'s work, we think that horizontal body locations can be used to anchor the display of personal information, which is one way to automatically manage screen space. This is one condition in our large screen real-estate management study described in Chapter 4. In Vogel et al.'s work users are able to browse public information on the rest of screen by reaching out their hands or shifting their body. This is useful for public displays, and might be used for in-home screens if a notice for all family members is available. But whether it will disrupt others needs further study. While proximity information is useful for transition between interactive phases for a public display, in a living room with limited size, interaction distance may not vary much, therefore proximity is not a major factor in this thesis. As our result indicates that users might not want their contents to be shown the moment they enter the room, the proximity factor could help determine when to display users' information in future studies. Our vision system can also provide gaze orientation information suggesting attention, and this information can be further studied. Although providing higher accuracy, the Vicon system requires users to wear markers which are cumbersome for interaction. The vision-based tracking system implemented in this thesis can provide both identity and user position information without asking the users to wear equipment.

Researchers have also studied location information of multiple users in smart interaction environments. These works are not confined to interaction with a display, and are often combined with other modalities such as speech to help issue commands and clarify user intention, as discussed below.

Aging-in-place research (Williams et al. [41]) is one example of where a smart environment based on semantic interpretation of video information is found. Williams et al.'s work presents the design and implementation of a low-energy multi-camera network that detects falling actions in an elderly living environment. The system uses the microcontrollers of Cyclops cameras to perform fall detection by background subtraction and calculating the aspect ratio of a detected person. The falling person is localized by successive pairwise image homography that maps the local coordinates of the camera that detects the falling person to a leader camera calibrated in the world coordinate. The accuracy of the falling events is 95% in a 40-image test set. The overall localization error is 15-22 inches for a 3-hop route. The goal of the system is to provide energy efficient fall detection camera network, so the capability of tracking is compromised: only one person can be detected at a time, and no identity information of the person can be provided. However, this is sufficient for their system, since the detection of fall events and localization in real-world coordinates is the priority. In our study the identities of multiple users are required, but only relative location is needed for interactive use, so we adopted algorithms that fits this need.

Bernardin et al. [8] developed a system for identifying and tracking multiple users in a smart meeting environment. The system fuses information from several fixed cameras including a ceiling mounted one into a parti-

cle filter framework to track multiple occupants. A set of steerable fuzzy-controlled pan-tilt-zoom cameras serves to opportunistically capture facial close-ups of people for identification. Voice as an identity cue is fused in to enhance the confidence level of the system. The accuracy of the tracking component reaches 73% on a test set of 100-minute recordings, and a tracking precision of 15 cm.

This system integrates opportunistic ID (to perform identification whenever a face image or voice cue is available) in a continuous tracking component for simultaneous identification and tracking, which is similar to the basic approach of our tracking system. While the system may demonstrate higher fidelity for long-term tracking and larger number of people, it does not rely on pure vision, and requires complex setup that may not be suitable for the use of an in-home display system. As described in Chapter 3, we developed a vision system simple enough to work within the setup and accuracy constraints of the smart display system, and will fit on a large display.

The above work demonstrates the possibilities of extracting location information using (potential) vision techniques, and informs us of several applications using this information. However, Vogel et al.'s system did not implement a vision component based on a commonly used consumer-electronics quality camera; the vision part of the Aging-in-place system lacks the ability to identify the users; and Bernardin et al.'s system might be too complicated for home use.

2.1.2 Posture and Gesture Based Interaction

Vision-based interpretation for human behaviors, specifically gestural and postural, can provide richer interfaces for interaction with displays than location information alone. This kind of interaction is extensively implemented and tested in various types of applications, especially in games and virtual environments [20].

Pantic et al. put forward the specific scientific and technical challenges of making computers understand human behavior in a survey [32]. The scientific challenges include determining the number of types of behavioral channels and the fusion, as well as the context and dynamics of human behavior. The technical challenges involve initialization, robustness, speed, training and validation processes. In our development of the hand gesture interpretation in Chapter 3, we use color pads to increase speed, simplify initialization and training, and improve robustness.

As a starting point of understanding human behavior, free-hand gestures have been widely studied. They have the advantages of natural and direct interaction, but intrinsic problems exist in gestural communication with computers [6], such as wrist and arm fatigue, the gestures recognized by the system not being easy for users to understand and memorize, lack of comfort because users are usually required to wear additional equipment, and issues of segmenting the gestures from movement signals. In Chapter 4 we design gestures suitable for the collaborative gaming context, and self-revealing enough for in-home casual interaction. We traded off some degree of comfort for improved accuracy by having the user wear a color pad, with

the expectation that if gestures do prove sufficiently useful, then more sophisticated hand gesture recognition systems will be developed to perform well with bare-hands in unconstrained environments.

Pointing is one of the most common forms of gesture envisioned for interaction in front of a large display. Static pointing, which is signaling certain objects to the smart system by pointing at them, is implemented by several systems. But in what context static pointing is useful is unclear because of the lack of user experience study involved.

Lee et al. [26] developed a vision system using two cameras coplanar with a computer monitor for pointing interaction. The system detects user's eye and finger tip locations by registering camera images with the flat panel display, and calculates the intersection of the eye-fingertip line with the display. This intersection point is used in an application to navigate target points in a menu. In the case of two cameras, the median angular error is less than 0.04 radian.

Demirdjian et al. [13] (Figure 2.1) developed an algorithm that supports understanding articulated arm motion using 3D cylindrical model of articulated appearance. The algorithm uses Iterative Closest Point (ICP) for 3D points alignment and a joint constraint reinforcement step to estimate articulated motion. They applied this articulated-body-based-pointer to the task of making the targets appear on a projected display. The average pointing accuracy (defined as the average of the distance between the target and the pointer) estimated during the task was about 20 pixels. This technology was applied in the MIT WebGalaxy [15] system for navigating hypertext information.

Figure 2.1: MIT Articulated Body Based Pointer by Demirdjian et al. [13] (©2002 IEEE).

Arm-Pointer [18] (Figure 2.2) is a real-world pointing interface. It detects real-world objects a user is pointing to. The direction and position of the pointing arm is made available by extracting people's shoulders and arms from images captured by stereo cameras combined by 3D calibration information. This interface can be applied to operate electric appliances at home. They report the best detection time of 2.74 seconds per object.

Many other systems map dynamic hand movements to certain commands to control the contents on the display. Dynamic gestures were used in early work to remotely control a TV. In Freeman's work [14] users move their hands to turn on the television and control changing the channel. On the display, a hand icon appears that follows the user's hand movement to help complete the task, as shown in Figure 2.3. Using hand image templates represented by orientation of image gradient, the system looks for the hand feature in the image and tracks hand movements. An open hand triggers

Figure 2.2: Arm-Pointer by Hosoya et al. [18] (©2004 Springer Science+Business Media).

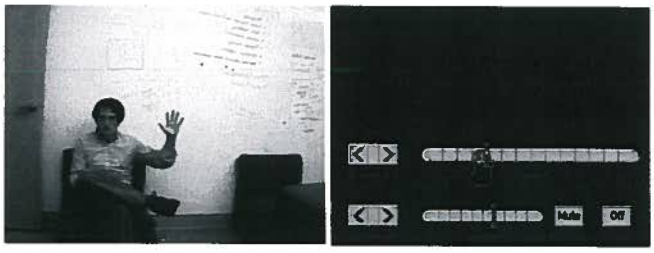

Figure 2.3: Television Channel Control by Hand Gestures by Freeman et al. [14] (©1995 IEEE).

the control, and left-right hand movements control up and down increments of channels and volume. Holding up a hand for a certain amount of time executes the commands. Freeman's work aims to solve the problem of lack of vocabulary for hand gesture control by providing visual feedback and simple mapping of hand movements to commands. However, it only supports movements in one dimension, and was not able to overcome some technical barriers such as small field of view of the camera (25°), and delay (half

a second). However, this work is one of the forerunners in this field, and has a lot in common with ours. They also observed similar phenomena such as user fatigue, a general problem with hand gesture interaction. We corroborate this finding and believe it has significant implications for what gestures are good for.

Figure 2.4: Ambient Gestures as a Controller for a Music Player by Karam et al. [22] (Copyright obtained from the authors).

Ambient Gestures [22] (Figure 2.4) is a vision-based gesture interface for a ubiquitous computing environment. It allows users to control music on a computer with hand gestures. Users are required to wear a bright color glove in the camera view and perform simple winding or horizontal/vertical hand movements to issue commands to a music application such as start, rewind, and select from a genre. Results of their informal user study show that the users were satisfied with this interface in spite of their initial suspicion. However, delay and recognition errors poses challenges to the user experience. The authors believe that the current non-visual interface is potentially ef-

fective for visual-display interaction in public and semi-public displays. We draw our idea of substituting hand detection with color pad detection from this work to provide higher fidelity hand tracking results for user study.

GWindows [42] is a hand gesture and speech combined interface that allows users to grab a window with their hands and move it across their desktop screen. Users can then perform actions such as to close, minimize, and maximize the windows. The system is composed of two parallel cameras, and assumes that the hands are the part in the image that is moving, and stays nearest to the cameras. The vision part of the system tracks the hands using multiple hypotheses including the moving part in the image, patch appearance similarity across adjacent frames, and binocular disparity. A preliminary, qualitative user study was conducted about using this interface. Users were generally impressed by the vision interface, but reported discomfort because of arm fatigue. This result is similar with the observation in our user study, and the fatigue problem seems to be even worse in the case of a large display. Their work also identifies possible extension of the interface, such as using hand gestures to complement keyboard and mouse when users are away from the keyboard, and if the display size gets bigger. This coincides with our consideration that gestures might be a complementary input modality for occasional use.

Krahnstoever et al. [25] (Figure 2.5) presents a framework for designing a multimodal natural interface with large screen displays. The visual components of this interface detect faces to initialize interaction, and track the head and the hands using skin color model and Kalman filter to prepare for the interaction. In the interaction phase, gesture recognition is

Figure 2.5: Interactive GIS Based on Gestures in Front of the Display by Krahnstoever et al. [25] (©2002 IEEE).

performed using Hidden Markov Models trained to recognize most natural gestures users would perform in front of a large display, such as pointing, and circling an area. Applications of this work include Multimodal Crisis Management System, retail and entertainment systems.

These four systems suggest in interacting with a smart environment, relative, dynamic hand movements could be more natural and expressive user interface than static pointing. They also provide properties of the interacting hands to facilitate detection, such as skin color cues, motion cues, and proximity to the display. Although they did not study the influence of complexity of gestures on user experience, we decided to use simple dynamic pointing gestures suitable for the game context in our user study.

While we learn a great deal from the vision-based interactive systems, these works have the following common insufficiencies for our needs:

Single user interaction

Most of them are not designed for multiple users. Although multi-user tracking has been realized using various approaches, the deployment of those algorithms for supporting interaction is rarely seen in literature, and often with the help of other modalities if there are any. In this thesis, we will look at the users' behavior in a system that tracks multiple people. A bigger technical challenge will be imposed on the system because of the greater uncertainty brought by multiple users. When it comes to interpreting user behavior, it is even harder to decide who is performing certain movements.

Static users

In most systems, especially gesture interaction systems, only static users are considered, e.g., in a standing or sitting pose. The users do not leave or re-enter the camera view in any cases. However, in a real world scenario, people come and go quite often, and coupled with the issues of multiple people, the dynamics in the process of interaction are complicated and not studied in these works. Our system in Chapter 3 addresses this challenge technically, and the first user study of our work detailed in Chapter 4 deals with the user experience issues regarding dynamic change of user scenes.

Lack of user experience and unclear user interface goals

Most systems were designed without a user survey or user experience study. This is partly because a vision system usually requires considerable training, tedious calibration and initialization process that makes it hard to be readily used. The limited effort in designing a vision-based interface to learn user behavior, and applying user behavior knowledge to the design of such systems has left us questions: are vision systems applicable, useful and

usable in a real world setting? If so, in what cases? Our work will try to answer the question from our experience of the user studies.

Interaction environment

Most systems are based on the desktop metaphor, as the active zone for the camera is relatively small, and the environment is more controllable. Hardly any systems are designed for large size screens. When switching from desktop display to large screens, usability issues such as real-estate management, sharing and collaborating, and fatigue in hand gesture interaction will arise.

2.2 Single Display Groupware Study

While most vision-based systems deal with single user interaction without study of user experience, we want to look at research on how multiple users interact with one computer display concurrently, known as Single Display Groupware (SDG). Although we ground our initial designs on this work, it has traditionally focused on design for work/meeting environments rather than in-home domains.

Systems in this genre deal with issues of screen access, and screen space management. Dynamo [19] is a large interactive display system that supports exchange of media in a communal space. It solves the problem of multiple user simultaneous interaction by providing each registered user a personal palette associated with their personal profile, along with a personal telepointer of the same color-coding. Multiple users can then transfer their media information from or to base interaction points such as USB disks and

MP3 players, or mobile interaction points such as laptops and PDAs. They are also allowed to display the media files on the common surface. Our work is inspired by the profile system provided in Dynamo, although Dynamo applies to a public setting and can be dynamically changed. In our work, we believe that family members will have their own, persistent profile requiring identification when they want to use the display.

Tse et al. [37] studied managing screen resources, particularly how users avoid interference in SDG by spatial and semantic separation of work. They found that spatial separation and partitioning occurred naturally across all participants, rarely requiring verbal negotiation. In their test, except for special task semantics, participants predominantly partitioned their workspace in a left/right fashion based on their seating positions, claiming their respective "side". In this thesis, we base our automatic placement of screens using the tracking information on this concept.

Further, Tsandilas and Balakrishnan [36] explored different techniques to reduce spatial interference not just spatial separation. In their work, three techniques are evaluated such that interference between the workspace of two concurrent users is minimized.

1. Shared screen: Users are allowed to utilize the entire screen, but only have access to objects that are owned by them or globally shared.

2. Split screen: Splitting the screen into one area per user ensures that interference between the workspaces cannot occur. Splitting can be initiated by a protocol or by the actual users. Panning or zooming can relax the problem of limited space.

3. Layers: Each user is provided with a layer of interaction. Each layer may be visible to multiple users, but its contents can only be manipulated by its owner.

Results show that the best approach in terms of performance is to share the entire display with appropriate use of transparency techniques for minimizing interference. Users also like to decide for themselves how they wish to partition the space, rather than pre-partitioning it for them. Compared to our work, their work assumes a working environment where users work together towards a common goal, so a lot of public information is needed. But at home, the user behavior is more heterogeneous, therefore they are more likely to interact with their own contents. Their work also assumes a desktop and mouse setting, where it is easier to manually manipulate screen spaces at will than to use a remote control for the large display. However, their work agrees with the result of our study that users need a certain level of control over the size or other properties of their spaces when managing the screen resources, as discussed in Chapter 4. User experience may be augmented when we combine automatic placement of users' space enabled by our tracking system and user-controlled expansion of individual spaces by overlapping others'.

Birnholtz et al. [9] discussed the relation between seating positions of users and the objects selected by them when engaging in a group task with the large display. They found that seating position has no influence on the horizontal coordinates of the articles selected to interact with. However, our results regarding the seating position and performance data provided in

Chapter 4 shows there is preference over having spaces laid out according to seating positions. This is because in their work, users are working in collaboration, and sitting in a hexagonal table much smaller than the display dimension at the center of the front of the display. The relations between display sizes and the preference over personal spaces placement needs to be further studied.

Another set of work studies the user input in SDG. The types of displays involve upright displays and tabletop ones. Hawkey et al. [17] studied the impact of proximity on the effectiveness and enjoyment of co-located collaboration on tasks on a single display. Proximity includes distance between the users, and the distance between the users and the display. Their results reveal interesting issues of device choices in interaction at a distance. They found that using a secondary display is problematic for collaboration for it compromises shared understanding and causes communication problems, while input at a distance is beneficial for collaboration. This finding indicates that using gesture control, which is a type of indirect input, can possibly benefit cooperative task on a large display.

Morris et al. [28] introduced cooperative gesturing, a type of interaction with touch sensitive input devices where multiple users gesture to issue a single, combined command. Their work summarizes the motivation for designing cooperative gesture interaction: increasing participation, drawing attention to important commands, enforcing implicit access control and providing entertainment.

SIDES [33] is a tool to help adolescents with Asperger's Syndrome practise effective group work skills using a four-player cooperative tabletop game.

The group work skills include negotiation, turn-taking, active listening, and perspective-taking for students in a social group therapy. The designers of SIDES use mechanisms to enforce turn-taking, hoping to give the players a feeling of ownership over the activity while still encouraging negotiation between the players.

The above two works are efforts to use computers for enforcing group work among team members in the case of SDG. They help us identify potential social benefits of designing a cooperative interaction environment based on user gesturing. Though implemented on tabletop displays, their designs shed light on the design of a collaborative game on large upright displays. We apply similar concept as in Cooperative Gesturing in designing the class of hybrid controller where gesture interaction complements the remote controller. When users have controllers of complementary functionality, they should have a better sense of cooperation because they rely on each other to complete the task, and no one dominates the game.

Birnholtz et al. [9] studied the influence of input configuration on group behavior in the collaborative task with a large display. Two conditions are involved in the study: a single, shared mouse, and one mouse per person. Groups of three users take part in a newspaper layout negotiation task. The group has a shared goal of laying out articles for the newspaper, and everyone has his individual goal of maximizing his points by selecting articles of certain topics he is responsible of. The study result shows that in the multiple mice condition, the groups work in parallel and work faster, and there is lower quality of discussion, while in the single mouse condition, the group take a team-based approach, and the quality of discussion ratings

improves, however it sometimes caused frustration to those not controlling the mouse, and provided the opportunity for one participant to dominate the task. Both Birnholtz et al.'s work and ours study the influence of input configuration on group behaviors: equal control and non-equal control, although our non-equal control configuration involves one high fidelity controller and two lower fidelity ones. Our results presented in Chapter 4 are similar to theirs in that the control is centralized in the non-equality group. However, not having individual goals make our results different with theirs in conflict levels; and complementary input devices differs our results from theirs in speed.

Another class of SDG work concerns the use of laser pointers in direct pointing to contents on large displays. Myers et al. [29] report that in interaction with a wall-size SMARTBoard, a laser pointer performs worst in speed among tapping directly on the SMARTBoard, using a PDA, and using a conventional mouse. Oh et al. [30] discuss several issues about using laser pointers for collaboration, such as spot detection using a camera, and difficulty of selection. They also present an evaluation of laser pointer performance compared to a mouse: the throughput of a laser pointer is 75% of that of a mouse. They put forward a method for identifying laser pointers of different users by powering each pointer in turn in a cyclic pattern. Vogt et al. [40] also present a technique to use a video camera to identify and track multiple laser pointers by blinking the laser to encode identities into an asynchronous serial bit stream. Laser pointer-based interaction is a subset of gesture interaction with handheld device, and has similar problems with barehand gestures in issuing complex commands such as selection. However,

as suggested in the abovementioned work, laser pointer interaction may require less computational power and is more accurate than barehand gestures, so might be useful in addition to the remote control of large displays.

The SDG work generally deals with interaction modalities of desktop display or large display with keyboard, mice, laser pointers and remote secondary display, or tabletop display. There is a lack of study on interaction of SDG using conventional controllers coupled with other types of input, such as free-hand gesture input.

2.3 Usefulness of a Vision-based System in User Interfaces

User experiences in vision-based systems have been studied. Karam et al. [23] explored user tolerance of recognition errors using hand gesture to interact with visual displays. Their work found that user tolerance is influenced by three major factors: user context, system performance, and user goals. The users are more tolerant of the recognition errors in a ubiquitous computing scenario than in a desktop scenario. The error rate can go up to 40% before the users abandon the gesture interaction in a ubiquitous computing environment. In terms of system performance, false positive errors should be eliminated from the system as much as possible. The user choices would change with the criticality of the tasks - if we change the criticality from simple timing to risks. In our vision system described and used in later chapters, we try to eliminate as many false positives as possible to minimize the fidelity effect on the users.

There are other works that discuss the difficulties of using hand gestures as an interacting modality in detail, such as in Kjeldsen's work [24]. He concludes that difficulties with responsiveness and accuracy means gesture interfaces are more appropriate for selecting and manipulating large on-screen objects. This conclusion justifies the role of hand gestures in our design of the hybrid device configuration: gestures are responsible for coarse-grain, large scale manipulations.

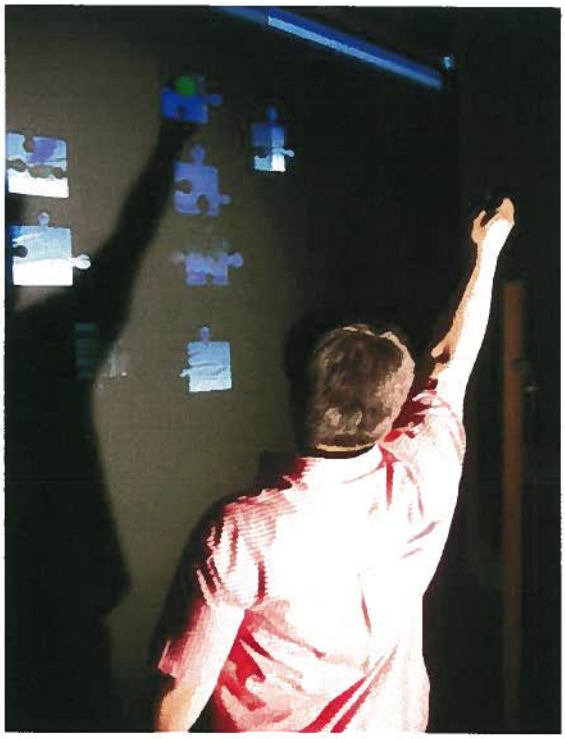

Figure 2.6: Shadow Reaching Prototype by Shoemaker et al. [35].

Shadow Reaching work [35] (Figure 2.6) explores the technique of using projective-based transformation of shadows for interaction with a large display. With the Shadow Reaching interface, users can have fluid access to

all areas of the display, and they can convey awareness of interactions to collaborators by changing their distance to the display. Shadow Reaching makes use of vision techniques to sense shadows, calculate user locations, and finally render virtual shadows on the display for interaction. The prototypes involve interacting with virtual balls and personalizing on-screen data using the shadow. This work also explores the design space of shadows using vision techniques, such as access control. The concept of vision-based whole body interaction is similar to ours, and they have demonstrated initial effort in exploiting the design space.

While accuracy is a big factor in the usefulness of a vision system in user interfaces with displays and is well understood, more factors, such as compatibility of types of vision techniques and type of interfaces need to be investigated. Shadow Reaching is one good initial work in this field. This thesis tries to contribute to this area by studying issues regarding usefulness of a multi-user tracking infrastructure for large screen space management and input device management.

2.4 Summary

Figure 2.7 summarizes the works discussed in this section categorized by the investigated issues in the thesis. Among the related works, most vision-based smart system works suffer from the drawback of single user interaction, static users, small display sizes, and lack of user experience studies. We developed a vision system that overcomes these drawbacks to a degree sufficient for a user study: it is able to handle multi-user identification and tracking tasks in

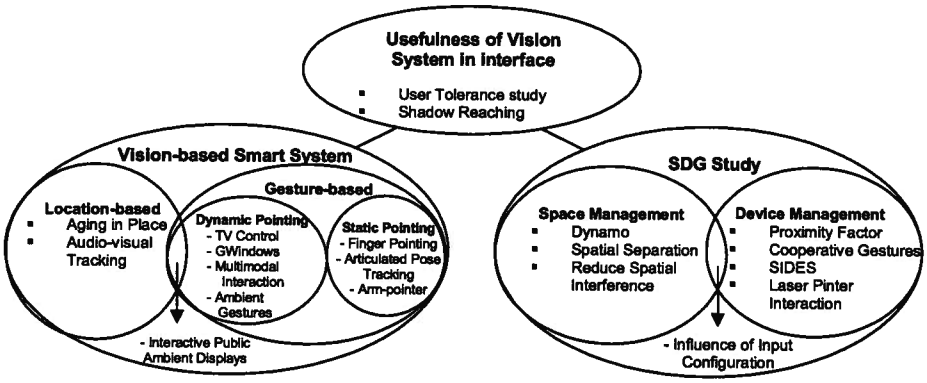

Figure 2.7: Summary of Related Work.

a relatively unconstrained, dynamic display environment. The SDG works we discussed inform us of multiple users' behavior in terms of managing screen space, and managing device combination when interacting with a single display. However, their works have the basic assumptions of a working or public environment, so whether their results could apply to an in-home large display environment requires further scrutiny. Our work extends the research on the two basic questions (managing space and managing devices) to a different interaction setting: interaction with a large in-home display using an automatic vision-based system. Finally, there is a considerable lack of work in discovering the usefulness of vision-based systems in user interface with the large displays, and exploring the design space. The major contribution of this thesis is the findings of whether and to what extent a vision-based system can help users manage their personal display contents easily, and use hand gestures to improve group collaboration.

Chapter 3

Vision-based Real-time Multi-user Identification and Tracking System

Identifying and tracking the locations of people has long been an interesting topic in the area of computer vision. These techniques are especially useful in the context of interaction with smart environments, providing knowledge of "Who are they?", "Where are they?", "What they are doing?" at least in a controlled indoor environment with a limited number of users. We can further analyse human behavior knowing the locations and identities of people.

As discussed in Chapter 1, vision systems can provide information to address some of the user interface issues, such as managing screen real-estate, and managing devices. We developed such a system as the basis for the user studies that investigate the two issues.

Note that there are many tracking techniques developed in the vision community, some more sophisticated and better in performance, yet we developed our own system sufficient for understanding the usability issues

of concern, instead of improving tracking itself.

In this chapter, we present our approach to enable real-time, multiple people identification and tracking in an indoor environment, and furthermore tracking of their hand movements to meet the requirement of an in-home interactive application. In addition, we describe real-time processing and communication infrastructure for interfacing the user study applications discussed in Chapter 4.

3.1 Problem and Proposed Approach

In this section, we abstract the multi-user identification and tracking problem in the context of interacting with large displays, and propose our solution using various vision based approaches.

Previous methods on Multiple Targets Tracking usually do not distinguish the identities of the subjects involved. Most methods deal with the problem of following the routes of different people whose identities are randomly initialized before each single test, such as tracking multiple hockey players [31]. However, an in-home application usually requires persistent identities of the users, such as for signing-in/out and placement of personal spaces in our screen real-estate management study. Luckily, the identities of users in the home are usually fixed: the members of the family, possibly the extended family and a limited number of friends. Therefore a tracking algorithm should know the locations of users appearing in the scene and their identities. As stated in the vision system challenges in Chapter 1, cameras mounted coplanar with a large display restrict their field of view; an in-home

interactive application should accommodate dynamic user setting and the environment, achieve sufficient accuracy and interactive rates with minimal manual calibration, adjustment or tuning.

Taking the above requirements into consideration, we summarize the multiple people identification and tracking problem as below:

- Registration and program initialization stage should be as simple as possible.
- Cameras are mounted coplanar with the display.
- The maximum number of people is fixed.
- The algorithm should make use of the characteristic information of people.
- The algorithm should adapt itself to the changes of people's appearance over time.
- The program should work with sufficient accuracy for the user study.
- The program should run with interactive rates.

To meet these requirements, we propose a tracking algorithm that fuses face information and clothing color information. Face detection and recognition is performed intermittently to identify users and update the clothing color template of a particular user, while color template based tracking is performed in parallel to keep track of each user in subsequent video frames regardless of whether faces are present. Users are able to update their color information voluntarily by showing their faces to the cameras. In this way,

a fixed set of people can be identified and tracked simultaneous, even if they leave the scene and come back in different clothing. The basic concept of fusing face-based identity information into color-based tracking is similar to Bernardin et al.'s work [8], but they implemented a system with more sophisticated camera setup (including ceiling-mounted ones) and audio cue.

3.2 System Overview

The proposed human tracking method consists of three parts: face detection and recognition, user detection, and user tracking. We also define two types of templates: *face templates* which are the characteristic human face patches for all users and are fixed; and *object color templates* keeping users' color histograms that are updated dynamically, and allows short-term appearance changes.

Figure 3.1 shows the diagram of this scheme. The face detection and recognition part (Section 3.3) detects faces in video frames and identifies the faces by comparing them with the face template. At the same time, the user detection part including background subtraction and bounding box detection (Section 3.4) detects bounding boxes of potential people in the video frames, without knowing their identities. Once a face is identified as a user, we pair it with the nearest bounding box, and use the histogram within this bounding box to update the object color template of this particular user. The tracking part (Section 3.5) then matches detected bounding boxes with object color templates, and labels each matched bounding box with the identity of its corresponding object color template.

The face detection and recognition part works in parallel with the user detection part. Results from both parts are fused in the tracking part to keep track of multiple people. We discuss the three parts in detail in the following sections.

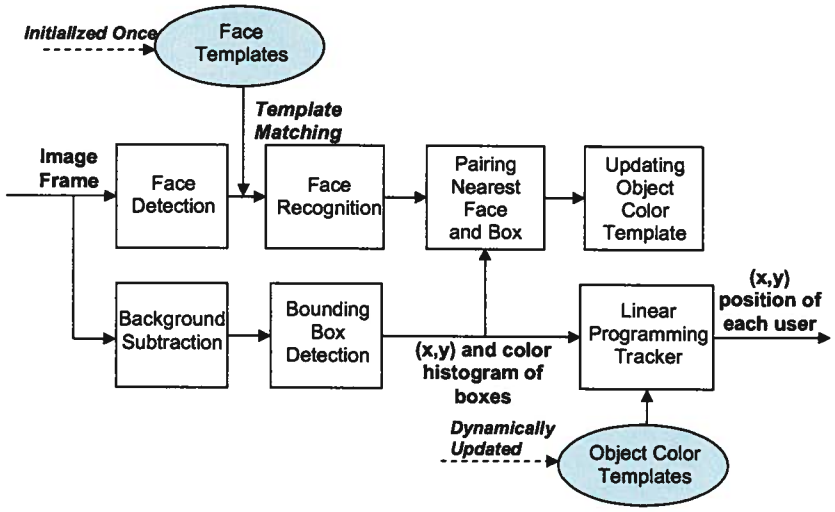

Figure 3.1: System Diagram. Face detection and recognition is used to update object color template. Bounding box detection is perfprmed in parallel. Tracking is done by matching detected bounding boxes with object color templates.

3.3 Face Detection and Recognition

In this section, we describe face detection and recognition techniques used in the tracking scheme.

3.3.1 Face Detection Using Boosted Face Detector

We use the boosted face detector [38] that is able to efficiently and reliably detect frontal upright faces in a controlled indoor environment. This method uses Haar-like features and a cascade of boosted tree classifiers. Since the face detection will be used for object color template update introduced shortly, we set a high detection threshold to keep the false positive rates low. Detected face patches are fed into a face recognizer to be assigned an identity or rejected if the face is not in the face template database.

3.3.2 Constructing Face Templates

Before the experiment, we ask each subject to stand in front of the camera and take a series of 20 grayscale images, and run the images through the face detection algorithm to extract 20 rectangular *FacePatches*. Then we apply the following method to each *FacePatch* for alignment. $FaceWidth$ and $FaceHeight$ denotes width and height of an extracted *FacePatch*.

1. Trimming. If $FaceWidth/FaceHeight < 3:4$, then keep the $FaceWidth$, and modify $FaceHeight = FaceWidth/3 \times 4$. The face patch is trimmed in height keeping $FaceHeight$ pixels that are vertically symmetrical to the center of the the patch. Else if $FaceWidth/FaceHeight \geq 3:4$, then keep the $FaceHeight$, and modify $FaceWidth = FaceHeight/4 \times 3$. The face patch is trimmed in width the same way as height.

2. Resizing. When $FaceWidth/FaceHeight = 3:4$, we scale the patch to a patch of 60 pixels wide and 80 pixels high using linear sampling.

3. Normalizing. We compute the mean value of all the pixel values of the face patch, and divide each pixel value by this mean. This step adjusts the template against indoor lighting changes.

Figure 3.2 illustrates the above steps. A detected face image is automatically trimmed, resized to a given scale, and normalized in pixel values. Then we

Figure 3.2: Trimming, Resizing and Normalizing a Detected Face Image.

take the average of all resized and normalized *FacePatches* of a person to construct his face template. The idea of recognition by taking the average of the training images is found in [11], which states that the face templates that are the result of averaging training face images produces far better results than a single instance, because it might coincide with human brain process of averaging mental images of human faces before recognizing a person.

3.3.3 Face Recognition

There are a number of ways to recognize human faces in new input images based on stored templates. Common practice includes simplistic template matching, which is to compute the Sum of Absolute Difference (SAD)(or the L^1 Norm) between each individual pixel value of a detected grayscale face patch and a face template. Discrete Cosine Transform (DCT) based recognition compares part of the DCT coefficients of the newly detected face

against that of a face template. Eigenface based on PCA is also a popular approach. In our work, we have tried all the above three methods, together with the Sum of Squared Difference (SSD)(or the square of L^2 Norm) instead of SAD. We found that in practice, when the resolution is not high enough or the person is not necessary standing near the cameras, the first simplistic template matching method exhibits the lowest false positive rate. Since the detected faces are all frontal upright ones, we do not expect the face recognition algorithm to account for a very big pose difference. Also, SSD seems to be sensitive to big random differences in a few pixels rather than the difference of overall features. Therefore, we use SAD as the face recognition approach in a natural setting.

When a new face is detected, it is first trimmed and resized to a patch of 60 pixels wide and 80 pixels high and normalized using the same steps as for face template construction. Then we recognize a face by computing the Sum of Absolute Difference between the face patch and each face template:

$$ID(NewFace) = \underset{k}{\operatorname{argmin}} \sum_{i=1}^{FaceHeight} \sum_{j=1}^{FaceWidth} |(NewFace(i,j) - Template_k(i,j))| \tag{3.1}$$

$Template_k$ is the face template of user k, $NewFace$ is the newly detected and normalized face patch. The ID k that makes $Template_k$ minimize SAD (smaller than a threshold) is assigned to the face. Thus we say the new face is recognized.

3.4 User Detection

Because faces are not visible or detectable at all times, we need other clues to keep track of people. In our work we use a person's full-body color histogram.

3.4.1 Background Subtraction and Bounding Box Extraction

To obtain the histogram information, we need to locate potential objects. We first do background subtraction and compute the binarized absolute value difference between an image and a background image captured beforehand. Then we apply sliding windows in different scales to the binary image with the pixels summed in each window. The sliding window roughly has a width-height ratio of a standing person (we assume the ratio is 1:3), and the pixel sum in the window is processed with

$$f(s, \mathbf{l}) = w(s, \mathbf{l})/A(s) \tag{3.2}$$

where s is the scale and $\mathbf{l}$ is the location of a sliding window; w is the sliding window pixel sum and $A(s)$ is the area of the sliding window at scale s;

We extract all the local maxima of $f(s, \mathbf{l})$ that exceed a threshold to determine all the potential bounding boxes of users. The procedure of extracting the bounding box of a user is shown in Figure 3.3. We further shrink the detected bounding boxes to tightly fit the binary foreground object. This step allows us to detect objects that are not in a standing pose, i.e., sitting and kneeling etc.

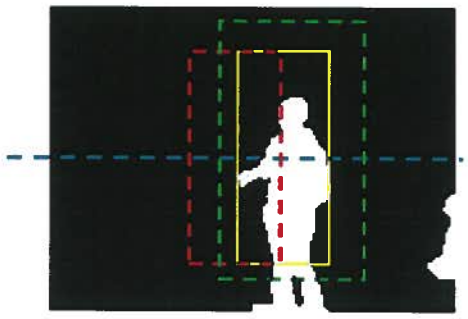

Figure 3.3: Computing $f(s,l)$ (density) value of sliding windows of different sizes, and extracting the local minimum to locate a potential user. The yellow window has a larger f value than the green window and the red window, so it is more likely to be the bounding box of a user.

3.4.2 Color Histogram Computing

We use color histogram in a bounding box as a second feature for representing appearance of a potential user. We quantify R,G,B color channels into 24, 24 and 16 color levels respectively, and join these values into a feature vector of 64 elements. We set a smaller number to the blue channel levels because the blue channel for a camera is the darker, less sensitive one, and usually contains higher level of noise. The vector is then normalized.

3.5 Tracking Solved as a Labeling Problem

Now that we have means to obtain the characteristic features of human objects - the faces, and the transient features, namely the color histogram, we propose a method to track human objects by integrating these two types of features.

3.5.1 Building and Updating Object Color Templates

We use face features to update the object templates, which are used in subsequent frames to identify the bounding boxes when the faces are not present, described as below:

- The template for each individual is a 64 element vector that represents the color distribution of a standard bounding box of the person.
- The templates are initialized as a value impossible for the appearance of a person in normal lighting conditions (such as all black).
- Once a face is detected and recognized, we take the bounding box closest to the face in horizontal coordinate as the one associated with the person that has the face, and update the current template of this person with the color histogram vector of this bounding box.

Figure 3.4 shows examples of face recognition and object color template construction. The green circles around the users' faces indicate that a face is recognized as a registered user. Whenever face recognition occurs, the histogram of the nearest bounding box (the red rectangular box around a person) is used to update the object color template associated with the user.

3.5.2 Matching Bounding Boxes with Templates

At the same time, detected bounding boxes are stored in a pool of potential users. Up to this point, tracking can be solved as a problem of assigning identities to the bounding boxes by comparing their histogram similarity with the object color templates. As this is a labeling problem, we compare

Figure 3.4: Examples of Face Recognition. The green circles represent recognition of faces.

two approaches to solve it: a Greedy approach and a Linear Programming approach [21]. In my experiment, I choose the Linear Programming approach for its ability to make use of temporal continuity.

Greedy Approach

The greedy approach pairs up the bounding box and the object color template that has the closest color similarity in each iteration, and marks any bounding box close enough to the paired box as being occluded. The matched box, the template and the occluded box are eliminated before the next interaction. The algorithm stops when all boxes are matched or marked as occluded.

Figure 3.5 shows a simple example. 1, 2, 3 are object color template labels and a, b, c, d, e are bounding boxes detected in current video frame. We compute color histogram differences between each bounding box and

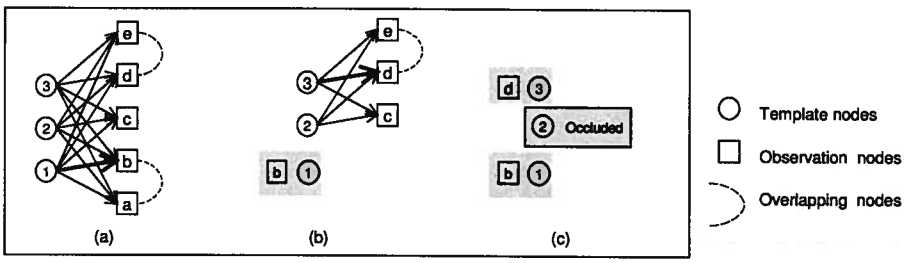

Figure 3.5: Greedy Approach for Multiple Objects Tracking. (a) Step 1. (b) Step 2. (c) Step 3. Circles are object color templates, squares are detected bounding boxes, and dotted arcs link overlapped nodes. Red arrows mean best matches between an object color template and a bounding box in each step. For example, in Step 1, bounding box b is matched with template 1.

each object color template. In this example, the pair $(1, b)$ has the smallest difference and is below the given threshold θ. We thus label bounding box b as user 1. Bounding box a is closely overlapped with b, which cannot be assigned a user label based on our occlusion constraint. a is then removed from the bounding box set. Template 1, and b are also removed from the graph. In the second iteration, template 3 and target image in bounding box d are found to have the smallest color histogram difference, and is a true match whose difference is smaller than the threshold θ. Similarly, we can remove bounding box e because it is closely overlapped with bounding box d. In this example, user label 2 is not assigned to any bounding boxes because its color histogram difference with the image patch in bounding box c exceeds threshold θ. Thus, in this example, person 2 is either not in the scene or occluded. This algorithm achieves a complexity of $O((n^2k))$, where n is the number of users in the template, k is the number of bounding boxes detected in a frame.

Linear Programming Approach

Despite the simplicity of implementation and the running time advantage, the evident shortcoming of the Greedy approach is that it does not make use of the temporal continuity. When the histogram of the users look alike, or the color values gather at high values, the greedy algorithm is error prone, thus the identities for the bounding boxes can be flipping around across frames. Therefore, we applied a more sophisticated Linear Programming approach (Jiang et al. [21]) to the matching of bounding boxes and object color templates.

The Linear Programming method models tracking as a multi-path searching problem. In this problem, each of the users has a path through his individual trellis graph. Each layer in the graph is modeled as one in a window of frames in the video. The nodes in vertical each layer are the observations of bounding boxes in each frame. The edges of the graph are modelled with a cost equation as below:

Cost = histogram difference between bounding boxes and each object color template

+ histogram difference between bounding boxes of adjacent frames

+ spatial distance of a particular user's bounding boxes in adjacent frames

All users' paths through their individual graphs can be optimized simultaneously using the total cost through their own graphs. The graphs are shown in Figure 3.6. Figure 3.7 shows an example of Amy's path and Gavin's path through a window of five frames. It is more efficient than ap-

proaches such as Dynamic Programming [7], and can achieve complexity of $O(n^3m)$, in which n is the number of users and m is the number of frames in each optimization window.

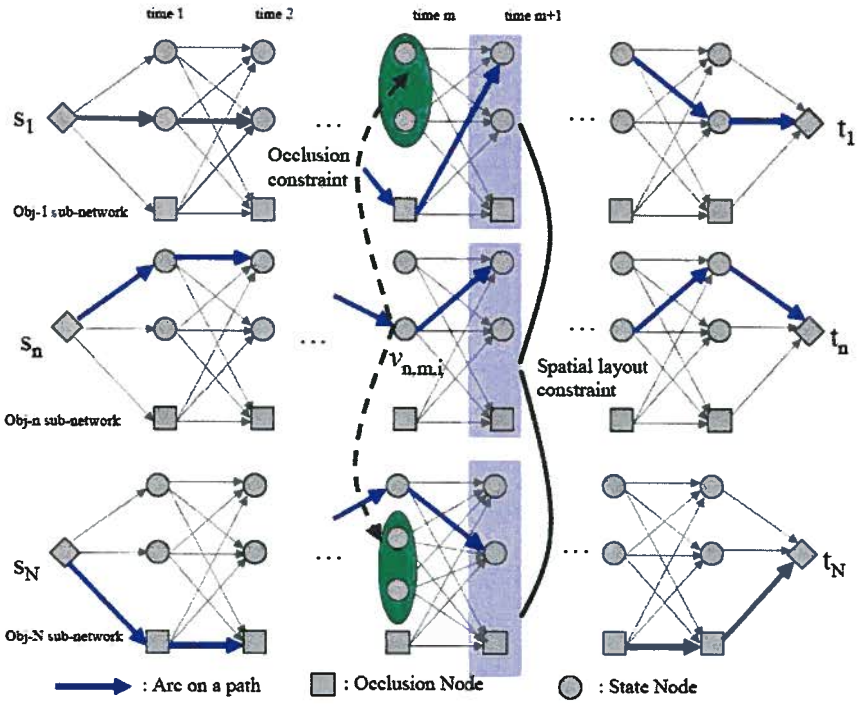

Figure 3.6: Linear Programming Approach for Multiple Objects Tracking by (Jiang et al. [21])(©2007 IEEE). S_1 to S_N are individual users' graphs. Each vertical layer of nodes represent one frame in the video. The bold arrows in blue in each user's graph represent his path in terms of bounding boxes through a window of frames.

Experimental Results of Multi-user Tracking

We implemented the multiple user tracking algorithm on a 2.6GHz PC running the Fedora 5 operating system. The incoming video is MJPEG format

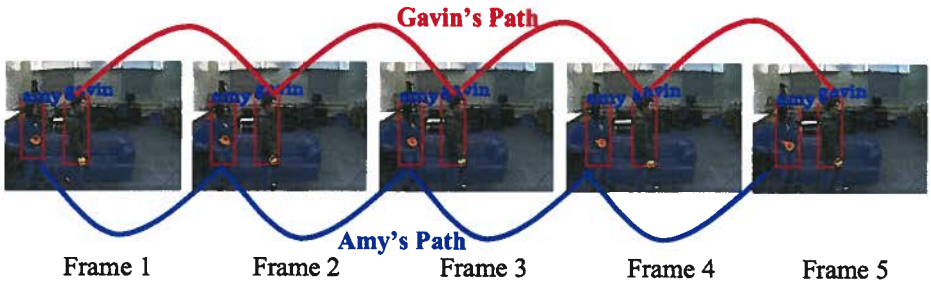

Figure 3.7: Linear Programming Approach Example. Gavin's path over five frames is shown in red, and Amy's path over five frames is shown in blue. Their paths are optimized through the frames using the cost function.

from AXIS 206 cameras. We used a camera API that decodes 320 by 240 JPEG images on the fly at 30 frames per second. In order to achieve interactive rates, we make the face recognition active every 5 frames, while other parts of the algorithm runs at 30 frames per second. The temporal window used in Linear Programming approach is 5 frames in our work.

We tested the approaches on two test sets and ran it against a ground truth file from manually marking locations of people in each frame. Table 3.2 shows the quantitative metrics for performance comparison.

	Frame Index	ObjectIndex	Horizontal Location
Ground Truth	i	j	x_{ij}
Test Result	i	j	(x^l_{ij}, x^r_{ij})

Table 3.1: Notation for Fidelity Metrics

Where:
$x_{ij}/x^l_{ij}/x^r_{ij} = 0$: The jth object in ith frame is occluded, or not in the camera view.
x_{ij}: x coordinate of the center of the human object in ground truth
x^l_{ij}: x coordinate of the left edge of the object bounding box in test result
x^r_{ij}: x coordinate of the right edge of the object bounding box in test result

Fidelity Measure	Implementation	Percentage
total number of appearance in ground truth	N($x_{ij} \neq 0$)	1
number of hits	N($x_{ij} \neq 0$ AND $x_{ij}^l \leq x_{ij} \leq x_{ij}^r$)	number of hits / total number of appearance
number of false positives	N($x_{ij} < x_{ij}^l$ OR $x_{ij} > x_{ij}^r$)	number of false positives / total number of appearance
number of misses	N($x_{ij} = 0$ AND $x_{ij}^l \neq 0$)	number of misses / total number of appearance

Table 3.2: Fidelity Metrics

Where
N(condition): Number of times the condition is true.
Number of appearance of objects: e.g., Amy and Sid are in frame 1; Amy and Joe are in frame 2; total number of appearance is 4.
Number of hits: Number of objects that are assigned correct identities
Number of false positives: Mis-matches of identities of appearing objects or taking someone is in the view while he is not.
Number of misses: The object is in the view but not tracked.

The experimental results are shown in Table 3.3. Figure 3.8 shows example frames from the tracking result. The red bounding box along with the names indicates that a user is identified and being tracked in the living room.

As mentioned, false positives are more serious problems than misses in the user study. We adjust the parameters of the system to minimize the false positives even more than these results. We believe that this fidelity is sufficient in the user study.

	Test Set 1	Test Set 2
number of frames	1593	720
number of objects	3	2
number of appearance	2939	1363
number of hits / percentage	2224 / 75.67%	1234 / 90.54%
number of false-positives / percentage	118 / 4.01%	50 / 3.67%
number of misses / percentage	597 / 20.31%	79 / 5.80%
Average 1-frame lapse	20.128 ms	20.163 ms

Table 3.3: Experiment Results for Test Set 1 and Test Set 2

3.6 Context-aware Data From the Tracking Infrastructure

Once the identities and locations of multiple users are known, we can extract other useful information based on this infrastructure. The extracted information can be hand locations of users and the users' distance to the large display. Determining distances requires at least two cameras. We use one camera in this thesis, so the distance information can be explored in future work.

3.6.1 Hand Tracking

The ability to track human hands may be a useful future technology in the area of human computer interaction. Hand tracking provides raw data for gesture based interfaces, which allows people to naturally perform manipulation to the computers, such as manipulating on-screen objects. The second investigation of our work is to examine the use of hybrid control devices including gesture-based ones, so we developed hand tracking techniques to

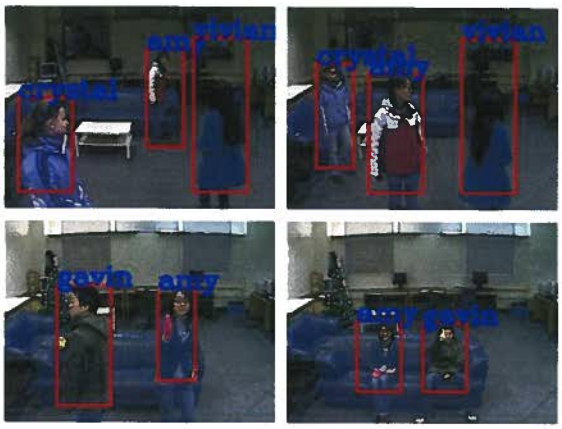

Figure 3.8: Selected Frames From the Result of Multi-user Tracking Program. The red bounding box along with the names indicates that a user is identified and being tracked in the living room.

enable simple gesture interfaces.

Difficulty of Tracking Human Hands

There has been rapid development on detecting and tracking human hands. However, most sophisticated techniques of hand tracking depend highly on the resolution of images, assisted by multiview data and simple color composition in the images, such as in GWindows [42]. Hand detection in natural, low resolution image settings still remains unsolved. Conventional ways of conducting hand detection are the combination of the skin color cue and the distance cue. As hands are basically a blob of skin colors, and are mostly used for pointing in interaction, we can determine hand areas using the following procedure:

1. Collect skin color image samples.

2. Split the images into H,S,V color planes, and ignore the V value to minimize the effect of lighting. Each skin pixel is represented as a vector:

   $\vec{x} = (h, s)$

3. Compute the mean $\vec{m}$ and covariance matrix C of all the sample pixels.

4. For each pixel within the detected bounding box, compute the Mahalanobis distance from the sample mean:

   $Dis = \sqrt{(\vec{x} - \vec{m})^T) C^{-1} (\vec{x} - \vec{m})}$

5. Set a certain threshold to the distance to extract skin color pixels.

6. Find connected components of all skin pixels, and consider the second largest connected area as the hand area (second to the face area).

The above procedure seems to be useful, however in interaction with a large display, the hands are usually occluding their faces. There are two potential ways to solve this problem. First, use movement cues. When interacting with large screens using hands, hand movements have larger amplitude and frequency than with small displays. Second, use cues from other views. In depth images obtained from stereo cameras, an area with smaller depth inside the bounding box is supposedly the hand area. An additional camera mounted on the ceiling might help, too.

Using Color Markers as Substitute for Hands

Because of the difficulties of tracking hands using skin colors in an unconstrained environment subjected to performance requirements, we applied

color based tracking, i.e. to follow predetermined color spots instead of skin colors. In the device management study, we asked participants to wear markers of different colors that are distinct from environment color.

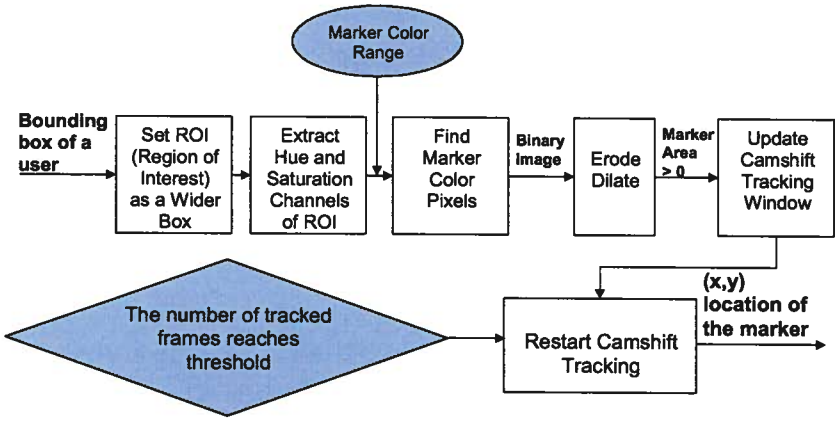

Figure 3.9: Hand Tracking Diagram.

We used the following Adaptive Camshift approach similar to one implemented in Liu et al.'s work [27] as the way to track users' hands. Figure 3.9 shows the diagram of this approach. We first store the color range of each user's distinct marker. Whenever a user is identified and being tracked, we widen the bounding box around him and set it as the "active hand region" or "Region Of Interest" (ROI). Then we split the image into Hue, Saturation and Value channels and find pixels within the range of certain Hue and Saturation values associated with the user, which are stored in advance. The output of this operation is a binary image in which potential marker color pixels are marked as 1 in ROI. We then perform Erosion and Dilation operations [12] to the binary image to extract the blob of all 1's at the location of the marker. Then we used Camshift algorithm [10] to start tracking this

	Test Set 2
number of frames	720
number of objects	2
number of hand appearance	1251
number of hits / percentage	1115 / 89.13%
number of false-positives / percentage	80 / 6.39%
number of misses / percentage	56 / 4.48%
Average 1-frame latency	2 ms

Table 3.4: Experiment Results of Hand Tracking

blob. At the same time, we continue looking for such color blobs. Whenever a blob is found, we restart the Camshift tracking at the blob location. This approach takes advantage of the continuity of the Camshift algorithm, and the accuracy and high false-alarm rate of color detection under varying lighting conditions and glove shapes, while avoiding the drifting problem of Camshift tracking. The center of gravity of the blob is considered the hand location of the user. We assign difference colors to multiple users so their hands can be tracked in parallel using the same approach.

Experimental Results of Hand Tracking

Figure 3.10 shows the example of tracking people and tracking hands at the same time. The results according to the performance metrics of tracking are shown in Table 3.4. We believe that this accuracy is sufficient for our later user study.

Figure 3.10: Example Hand Tracking Result. The orange circles represent the detected hand locations.

3.7 Real-time Processing

This section describes the infrastructure for capturing and processing images to achieve interactive rates for the user study.

3.7.1 Overview of the Real-time System

We aimed to build the vision-based system as close to real-time as possible in order to achieve the performance sufficient for interaction. The program flow of this infrastructure is shown in Figure 3.11. The data flow sequentially from the capture thread to the processing thread, and finally to the display thread. A pthread_join is used to synchronize the three threads. The three threads process at the same time, and when all of them finish processing, another loop starts again. Therefore, to improve the performance, we not only need to optimize each thread, but also balance the processing time in

the three threads, and at the same time keep the fidelity at a reasonable level. This vision-based infrastructure is implemented in the C language with the OpenCV library [2].

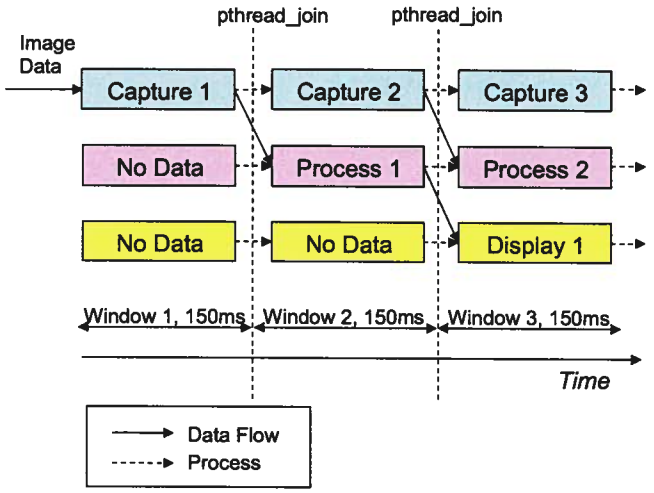

Figure 3.11: Flow Diagram of the Tracking Program.

3.7.2 Video Capture

The video captured by AXIS 206 Network Camera is done at 30 frames per second in MJEPG format. We use a decoder that can decode JPEG images in the memory [43]. There are two threads involved in the decoder. One thread constantly receives images streamed from the Network camera, while the other thread reads in the image and decode it. When the decoding is done, the second thread polls the first thread for the current image. In this way, there could be frame loss if the decoding rate is lower than the streaming rate, but less delay, which is what we hope for an interactive system.

3.7.3 Video Data Processing

The video capture thread sends a sequence of captured images to the processing thread for image processing. Since the Linear Programming algorithm performs optimization over a window of frames, it needs to wait for this window of frames to arrive before processing them as a whole. To balance the latency and the optimization result, we set the optimization window to 5, so it takes $30ms \times 5 = 150ms$ to capture the first set of frames. We then applied the following approaches to improve the processing performance to match that of video capture, so the next batch of frames can be processed directly after being captured:

- Reduce the resolution for background subtraction and bounding box extraction.

  Extracting bounding box from an image is a time-consuming part, because it searches for various sizes of bounding boxes for decision. However, as we only want to know roughly where the users are, we downsample the frames fed into the bounding box extractor from 320×240 to 160×120.

- Perform face detection every 5 frames instead of every frame.

  Searching for faces in an image is computationally costly as well. There are two ways to speed up face detection: reducing the detection rate, and reducing the resolution of detectable faces. Since users' full-body locations do not change much within 150ms, we do face detection and recognition at the rate of every 5 frames. However, we did not

reduce the resolution for face detection, as it is critical for the template update. We do restrict the range of the face size to look for. This step not only improves performance, but also avoids the situation when people are too far from the camera when their face image resolution is too low, or when they are too close so their face images tend to deform.

By doing the above improvement, the processing time for 5 frames with Linear Programming algorithm is reduced to lower than 120 ms - less than the capture. Together with the initial 150ms for capturing the first 5 frames, there is a 0.3s delay for interaction. Future improvement includes using a sliding window with incremental Linear Programming, namely recalculating after every frame comes in using a window size of 5 to keep the latency down. A latency of 0.3s is sufficient for the coarse-grain gesturing task as big hand movements cannot be completed very fast.

3.8 Communication Model

Our goal is to study several interaction issues using the tracking infrastructure. To do this, we need a model to combine the tracking processing units, the application, and various other devices.

We use a client/server architecture with all computational components wrapped in Python. The processing components (the tracking system) act as servers and communicate using TCP/IP to send Python commands to other client components (e.g. the user study application) coupled with data. Figure 3.12 shows the different components.

Figure 3.12: Communication Diagram. The tracking program acts as a server wrapped in Python, and the user study application is a client that receives commands and sends data to the server.

3.9 Summary

In this chapter, we described multi-user simultaneous identification and tracking algorithms, the real-time processing structure, and the communication model. We think that the fidelity of full-body tracking and hand tracking is high enough for the user study, and we have struck a proper balance between accuracy and speed to meet the requirement of the application to be used.

Chapter 4

Vision-based Multiple-user, Multi-device Interaction with Large Displays

According to the motivation and goals described in Chapter 1 regarding multi-user interaction with large displays, we want to see whether users prefer to manage their spaces on the large display using the vision-based system, or conventional remote controllers. We also want to look at whether low-fidelity gesture control can be a useful complement to high-fidelity remote controllers in a group collaboration context. In this chapter, we investigate these two issues through user studies.

The design of this chapter is shown in Figure 4.1. We designed two studies, one regarding large display screen real-estate management, and the other for device management. To explore usability issues in managing large screen real-estate, we designed two experiments to examine the sign-in/out process (Identity Experiment), and personal space placement (Placement Experiment) respectively. The design conditions and prototype applications are specified in the following sections, along with the study results and

discussion.

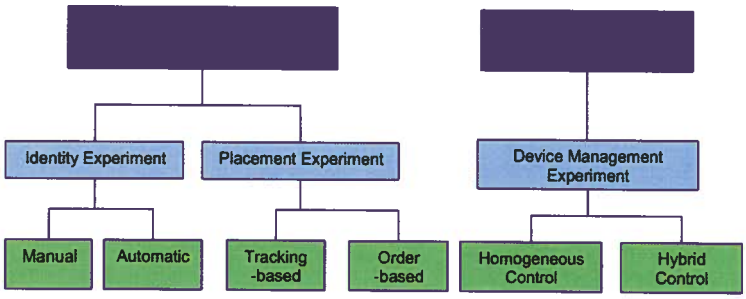

Figure 4.1: User Study Design.

4.1 Study One: Screen Real-estate Management for Large Display Interaction

As the tracking infrastructure is able to provide information about who is in front of the large display, and their relative locations, a smart large display can take advantage of this information to automate the activation and placement of multiple users' personal spaces, as was discussed in Chapter 1. We designed a user study to investigate whether tracking can be useful in automating the management of personal spaces from users' perspective.

Figure 4.2 illustrates the basic test scenario of this study. Three users sit in chairs facing the large display TV in the experiment room that simulates the "living room area". Each of them is given a mobile phone as the remote controller for changing and manipulating contents on the interactive large display TV. Each of them has a profile and personal contents (pictures in this case) stored in a server computer that controls the display contents on the TV. Everybody has his "personal space" on the TV, which is a window-

like area with his/her name on it. Users need to sign into their personal spaces when they want to start interacting with their personal contents, and sign out when they leave the living room area. When multiple users have signed in, the large display will be horizontally partitioned into several areas, each area holding one person's personal space. Users are asked to spot objects in a series of pictures shown in their personal spaces and to answer multiple choice questions in answer sheets as a form of interacting with their personal contents. They are prompted to leave the "living room" area occasionally during the task to simulate a dynamically changing scene. A network camera mounted on top of the large display captures the scenes in the living room and sends the video to a server computer for analysis.

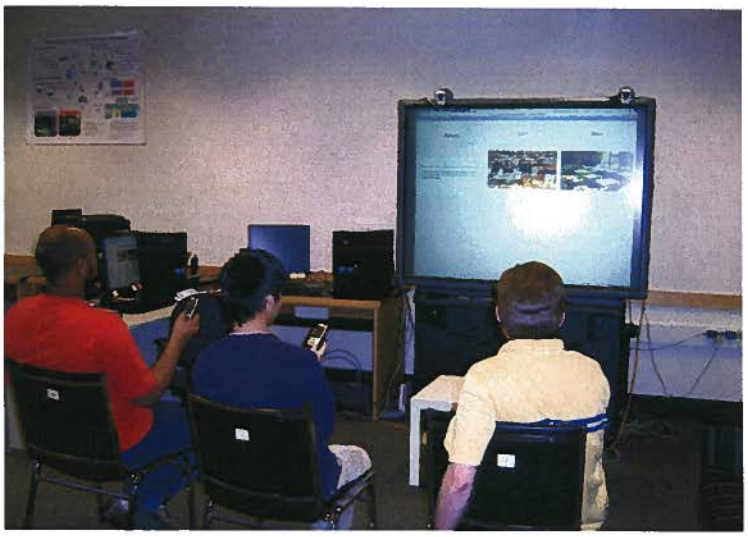

Figure 4.2: Screen Real-estate Management Study Setup.

There are two different issues involved in this test scenario, of which the first involves the means of signing-in/out of personal spaces. There are two

ways for registering personal spaces. One method assumes that each person has their own remote control, or that a common remote control has a way to identify the person holding it. Key presses on the remote controller signifies the intention of interaction. The personal space will pop up the moment the user presses a "sign-in" key, and will disappear once the user presses a "sign-out" key. The second way to register personal spaces is to use our vision infrastructure. Once the server computer running the vision program senses that someone is in the living room area (the camera view), it signals to the display and activates one's personal space on the display automatically.

The second issue arising from the test scenario involves the placement of users' personal spaces once they sign in. We partition the large display in a horizontal fashion placing one personal space in each partition as it coincides with users' natural tendency found in [37]. There are various approaches to ordering the spaces. We considered comparing two placement strategies: 1. a fixed, left-to-right order as each new person logs in and 2. relative to where the person sits when he logs in. With this latter approach, for example, the person who sits on the left will get his space placed on the left of the screen, so that the crossing of sightlines are minimized. The relative physical locations can be analysed by the vision program on the server computer based on the camera feed.

The following subsections describe the experiments we designed for each of the above issues.

4.1.1 Identification Experiment Design: Manual vs. Automatic Sign-in/out

This experiment deals with the first issue discussed above. It aims to find out whether users prefer to manually control the sign-in/out process using a conventional control, or have it automatically done by the smart vision system.

Hypothesis

The hypothesis of this experiment is:

Users prefer to be automatically signed-in and signed-out of his space rather than manually.

Independent Variables

The independent variable of this experiment is the sign-in/out mechanism. We compare two conditions of this variable to test the hypothesis.

Condition 1. (Manual sign-in/out) The users choose from a menu in a programmed mobile phone to sign in/out.

We enable the users to choose their names from a menu on mobile phones to sign in. When one person signs in, all other users' spaces gradually get resized, so that everyone gets the equal width for his space. When a person leaves the interactive area, he needs to use the menu to sign out. When a space is signed out, it gradually disappears, and all other existing spaces expand to share an equal width on the screen. Also, we offer users menu items to sign out other people who

have forgotten to sign out before they leave the interactive zone. At the same time, we make the interaction a bit more difficult using cropped pictures (images are cropped to 1/3 of its original width, shown in Figure 4.3) in personal spaces to encourage users to sign others out. If they do sign others out, it would imply that the automatic sign-out will be helpful. The personal spaces are placed according to their physical locations the first time they sign in. When users sign back in, their spaces will pop up in the same place as before.

Condition 2. (Automatic sign-in/out) The activation of users' spaces is automatically determined by whether they are in the interactive area with the help of the tracking infrastructure. Users do not need to sign in/out explicitly using the mobile phone.

Users need to show the camera their face to initialize the tracking process. But when they leave and subsequently return, their spaces will be shown according to their presence, and the spaces are aligned with their physical locations.

We will conduct a two-condition, within group study for this experiment.

4.1.2 Placement Experiment Design: Order-based vs. Tracking-based Space Placement

This experiment deals with the second issue. It attempts to learn whether order-based placement or tracker-based placement provides the users with lower level of distraction and better overall interaction experience.

(a) (b)

Figure 4.3: (a) Pictures without being cropped. (b) Pictures cropped.

Hypothesis

The hypothesis for this experiment is:

Aligning a user's workspace on the screen according to his relative horizontal physical location with other users helps to improve his task performance and overall interaction experience.

Independent Variables

The independent variable of this experiment is the mechanism for placing multiple users' spaces on the large display. We have two conditions to compare:

Condition 1. (Order-based) The horizontal placement of users' workspaces is determined by the order of signing-in.

When a user arrives at the interactive zone, he uses a simple key-press on the mobile phone to sign in, and will be allocated a space to the right of all the existing spaces. When a user signs out and signs back in, he will retain the spot he was originally occupying.

Condition 2. (Tracking-based)The horizontal placement of users' workspaces is determined by the relative physical locations of users, which is the output of our tracking infrastructure.

A user uses a simple key-press on the mobile phone to sign in, and will be allocated a space according to his relative physical location with other interacting users. Therefore, the spaces can get reordered if users switched relative locations, or if they sign in or out. At the beginning, the users are required to show their faces to the camera to initialize the tracking program.

We will conduct a two-condition, within group study for this experiment.

4.1.3 Apparatus

For both experiments, we set up the study system in our lab that simulates a living room. We used a 66" SMARTBoard 3000i to act as the interactive large TV display with two cameras mounted on top, though only one camera is used in this study. Chairs were laid out horizontally in front of the TV at a distance of 2.5 meters.

The remote control is implemented on the mobile phone Nokia N80. A Python script runs on the mobile phones sending key-press input to the study application. The key-press commands include sign-in/out, advancing pictures, going back to previous pictures, and panning the pictures in the Identity Experiment (Refer to Section 4.1.5). The mobile phone prototype is shown in Figure 4.4.

A 2.6GHz PC running the Fedora Core 5 operating system is the server

Figure 4.4: Mobile Phone Controller for Screen Real-estate Management. Subjects can sign-in/out and manipulate pictures on the display using the mobile phone.

who receives the camera feed and runs the vision algorithm. Using the communication model described in Chapter 3, the prototype application serves as a client running on a different PC that gets output from the vision algorithm in tracking-based conditions. The application itself acts as a server that waits for the commands from the mobile phones via the wireless channel.

4.1.4 Application

We developed a prototype application for this experiment using the Python programming language, as is shown in Figure 4.5. The application allows three people to have their personal spaces displayed on the large display in different orders according to the test conditions, and prompts people to leave the living room alone, in pairs, and in triplets.

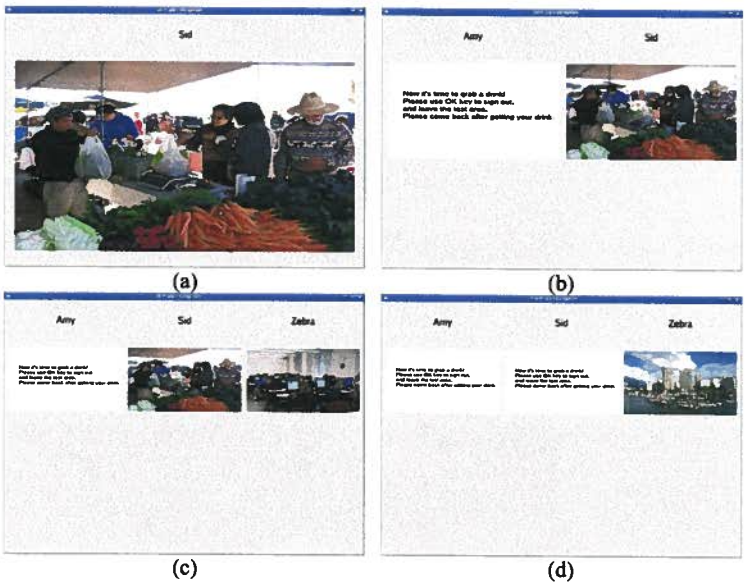

Figure 4.5: Application for Screen Real-estate Management. (a) One person using the display. (b) Two people share the display. (c) Three people share the display, with one person prompted to leave. (d) Three people share the display, with two people prompted to leave at the same time.

4.1.5 Task: Object Spotter

In both experiments, users are expected to spot objects in a sequence of 20 photos and answer multiple choice questions on the answer sheets as fast as possible, and as accurately as possible. Each photo corresponds to one group of five possible choices on the answer sheet. The subjects are supposed to place tick marks on the answer sheet next to all the items they see in the photo. Their spaces have their names and the picture numbers associated with them. They are occasionally prompted to leave the interactive area to check their answers, and come back in. We control the movements of subjects with appropriate timing of on-screen prompts so that each user

experience leaving the interactive area alone, in pairs and in triplets at least once in each session. This is to simulate the case of getting a drink or taking a break. Subjects are provided mobile phones as remote controllers. They can use simple key combinations or menus to sign in/out of their spaces, and advance/go back in the pictures during the task. In the Identity Experiment the pictures get cropped when more than one user shares the screen real-estate, showing only the central part, as shown in Figure 4.6(h). Cropping the pictures makes it difficult for other users should one person forget to sign out in the manual condition. It creates a bigger difference between the conditions. In this case, users are supposed to use the mobile phone keypad to pan the pictures to see the whole picture. In the Placement Experiment, whenever users' spaces get resized, the aspect ratio of the pictures is kept as 16:9. [1]

Figure 4.6 exemplifies one possible scenario: Bob signs in first and is allocated the full screen for his space (a). Later Jane signs in, and the screen splits in half for both of them to share the space (b). After a while, Jane is prompted to leave the room (c), so she signs out of her space and Bob's space expands to full screen again (d). Later, Mike signs in, then Bob and Mike share the screen (e). After Jane comes back, three of them share the screen (f). Occasionally, Bob and Mike are prompted to leave the room at the same time (g). (h) shows the cropped pictures for the Identity Experiment.

We hope to see that subjects spot objects faster and more accurately in

[1] We choose 16:9 as it is the ratio of popular wide-screen TVs. Some of the test pictures whose aspect ratio is not 16:9 were stretched to maintain the ratio of the personal spaces.

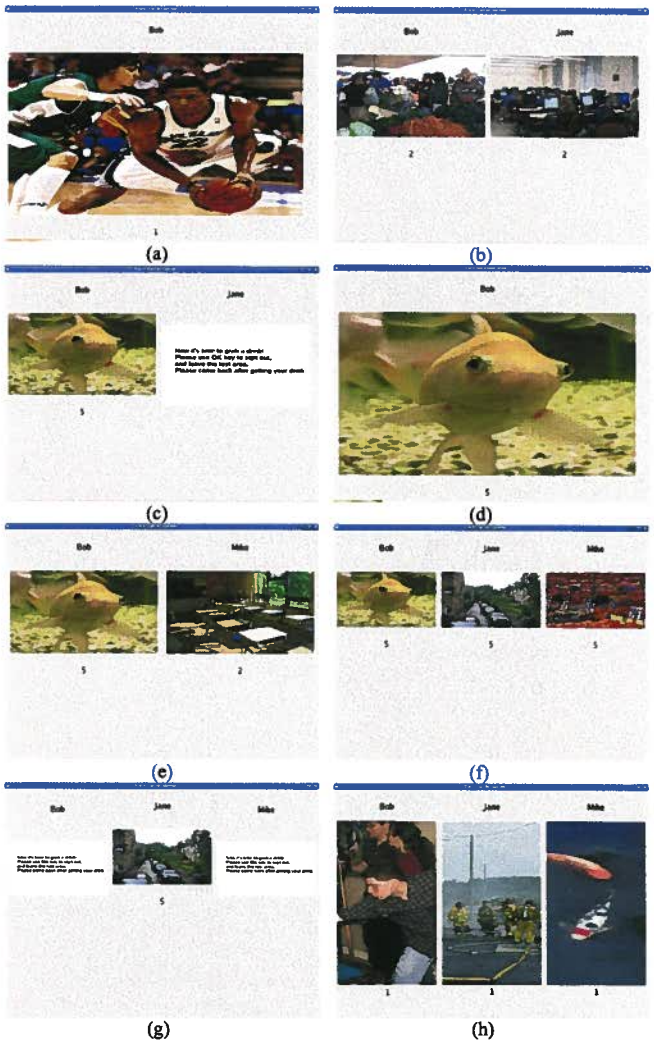

Figure 4.6: Examples of Screen Real-estate Management Study Experiment. (a) Bob signs in, and enjoys the full screen. (b) Jane signs in later, and shares the screen with Bob. (c) After a while, Jane gets the notice to leave the room. (d) Jane leaves the room, and Bob's space expands to full screen. (e) Mike signs in, and shares the screen with Bob. (f) When Jane comes back, three of them share the screen. (g) Bob and Mike are prompted to leave the room. (h) Cropped pictures in Identity Experiment.

automatic/tracking-based conditions in both experiments, as well as have a greater sense of ease in these two automatic conditions.

4.1.6 Participants

We recruited 13 subjects for our pilot studies. And for the formal study, we recruited 18 subjects, 6 male and 12 female. Sixteen of them are undergraduate and graduate students at the University of British Columbia, whose majors range from Science, Engineering, Arts to Business. Two of them are non-students. Their ages range from 20 to 59 years old. Most of the subjects use computers for more than 30 hours a week, and watch TV programs for less than 20 hours a week. Most of them hardly ever experience interacting with a large screen display, except for the playing of games once or twice such as Nitendo Wii at a party or alike. We divided the subjects into 6 groups of 3. All groups sat for both the Identity Experiment and the Placement Experiment, whereby each of them had to share the large display with 2 others in the group.

4.1.7 Procedure

We ran the study for a total of four different sessions, two sessions per group for the two conditions in Identity Experiment, and two sessions per group for the two conditions in Placement Experiment. Before the experiment, we offered a brief training session where subjects tried manipulating the picture sequence with their mobile phone controllers. The whole test lasted for about one hour. The order of sessions in each group is shown in Table 4.1.

Group Index	Session 1	Session 2	Session 3	Session 4
1	P1	P2	I1	I2
2	P2	P1	I2	I1
3	P1	P2	I1	I2
4	P2	P1	I2	I1
5	P1	P2	I1	I2
6	P2	P1	I2	I1

Table 4.1: Session Arrangement for Screen Real-estate Management

Where
P1: Placement Experiment(Order-based)
P2: Placement Experiment(Tracking-based)
I1: Identity Experiment(Manual)
I2: Identity Experiment(Automatic)

4.1.8 Measures

We measure the average time for spotting the objects in each picture, and the total mistakes they make in each session. These are indices of users' level of attention, disturbance and convenience. Users were also asked to fill in a questionnaire to rate and comment on their level of annoyance, speed, sense of control and general satisfaction when sharing the screen with their group members. The questions can be found in Appendix A.2.2 and A.2.3. All our rating data uses a 5 point Likert Scale where 1 represents Strong Disagreement and 5 represents Strong Agreement.

4.1.9 Experimental Results for Screen Real-estate Management

The results from the Identity Experiment indicate that users did not show significant difference in speed or accuracy between two conditions. While

they liked the idea of automatic sign-in/out, they preferred the manual control in the aspects of control and ownership. Results of the Placement Experiment demonstrate that while users did not show significant difference in speed and accuracy, their ratings and feedback suggest they preferred automatic placement in terms of its ease in interaction.

Results for Identity Experiment: Manual vs. Automatic Sign-in/out

Table 4.2 shows the average time spent on one picture and number of errors made out of 100 items for both conditions, and standard deviation. Subjects are 0.91 second (5%) faster and make 0.41(9.2%) fewer errors on average when they can be automatically signed in and out. A t-test shows that the difference in average speed is not statistically significant ($t(16) = 0.875, p = 0.395$); and the difference in average number of errors is not statistically significant either ($t(16) = 0.574, p = 0.574$).

	Speed		Number of Errors	
Placement Type	Mean (s)	SD	Mean	SD
Manual	19.33	4.49	4.88	2.26
Automatic	18.42	3.92	4.47	2.67

Table 4.2: Speed in Second and Number of Errors of Identity Experiment.

Questionnaire results suggest that while users liked the automatic sign in/out enabled by the tracking system, they also needed to maintain a certain level of control. Fourteen out of 18 subjects commented that the feature of automatic sign-in/out as "convenient", "interesting", "useful" or "cool". Only one person did not like looking at the camera to sign in because she felt

that similar to going through a security check. This is due to the fact that the current prototype made the camera very obvious and possibly imposing. Likely, building the camera into the design styling of the display might reduce this effect. Detailed user comments are provided in Appendix A.3.1. The subjects rated 4.13/5, 4.12/5, and 4.12/5 on average for the statements "I felt easy to log in because I only needed to show my face to the camera.", "I liked the functionality that my space is gone after I went out." and "I liked the fact that when someone went out, his space was gone by itself." respectively. However, subjects rated control and ownership higher in the manual sign-in/out condition, as shown in Figure 4.7. Ownership, in particular, had a mean rating of 4.35 for manual sign-in/out while it was only 3.59 for automatic sign-in/out showing a significant difference ($t(16) = 2.75$, $p = 0.014$). This finding indicates that although subjects found it interesting to be able to automatically sign in and out, they still liked to have a certain degree of control from a type of controller. One of the reasons might be the subjects observed and knew that the tracking infrastructure could make mistakes, and the errors could cause more serious problems or inconvenience than mere misplacement of the spaces. They might not be able to find their space when they were ready to interact, or someone else's space could still be there when he was gone.

Only 4 subjects out of 18 signed out other people, of which 2 did this to make room for the interacting subjects. The questionnaire comments revealed that one subject did this just to try out the feature, and another subject did this by accident. Others either did not want to do this because of social courtesy, or were too concentrated on their own work to do this.

Therefore we do not take this as an indication of the usefulness of automatic sign-out.

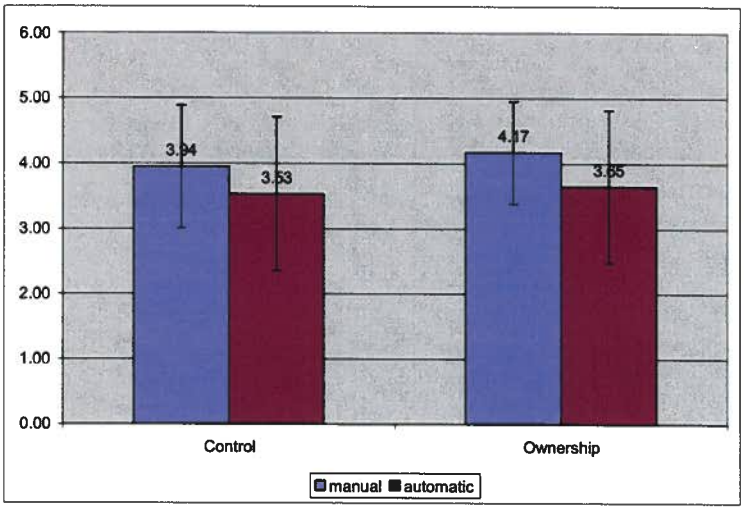

Figure 4.7: Comparison of User Ratings in Control and Ownership for Identity Experiment.

Results for Placement Experiment: Order-based vs. Tracking-based Space Placement

Table 4.3 shows the average time spent on each picture, and average errors made out of 100 items for the different placement conditions, along with the standard deviation. The result indicates that subjects are on average 1.1 seconds (9.4%) faster and make 0.84 (18.9%) fewer errors on average when using the tracking-based condition. However, a t-test shows that the difference in average speed is not statistically significant ($t(17) = 1.40, p = 0.18$); and the difference in average number of errors is not statistically significant either ($t(17) = 0.69, p = 0.5$).

	Speed		Number of Errors	
Placement Type	Mean (s)	SD	Mean	SD
Order-based	12.69	2.99	5.28	3.83
Tracking-based	11.59	1.69	4.44	3.17

Table 4.3: Speed in Second and Number of Errors of Placement Experiment.

Results from the questionnaire show that the subjects felt that the placement was natural and they had a greater sense of control and ownership when their spaces were aligned with their physical positions, albeit only slightly.

People did find that crossing other users' sightlines was annoying while doing the task, especially when the spaces were placed at the opposite end of the screen to where the person was sitting. Ten out of 18 subjects made comments such as "cannot see (my contents) clearly", "inconvenient", "difficult", "irritating", or "confusing" when having to look into the spaces not lined up in front of them. Five other people did not mind the crossing of sightlines, either because they thought they sat close enough to their spaces despite the mismatched relative locations, or they were too concentrated on the tasks to notice. Detailed user comments are provided in Appendix A.3.1. Subjects highly rated the tracking-based condition as reflected in the ratings for statements such as "quickly and easily identify my space", "easy to work in my space", "have enough control", and "feel that my space belongs to me". For the order-based condition, they rated highly "hard to resume work after coming back", as shown in Figure 4.8.

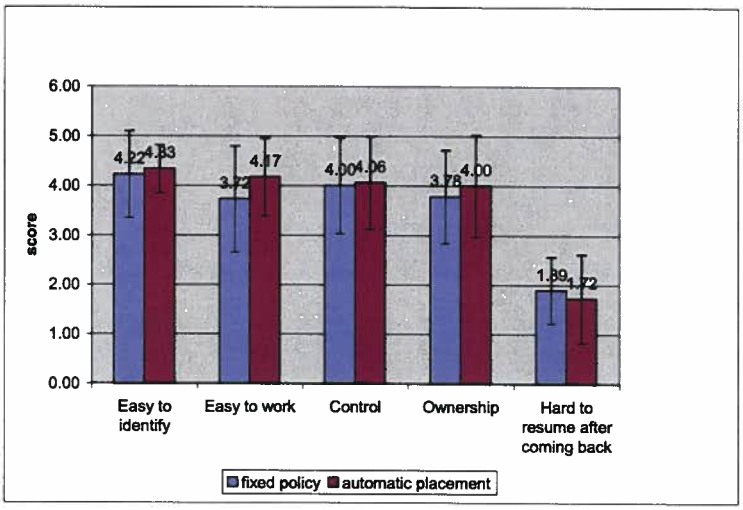

Figure 4.8: Comparison of User Ratings for Placement Experiment.

Other issues for Real-estate Sharing

We also conducted a survey to elicit the subjects' opinions about sharing spaces on their TV. Subjects think that the individual spaces are useful for gaming on the TV, but they would like to have their screen sizes static. When it comes to watching a movie or TV programs, some people prefer a full screen. However, it is not clear how this would scale to wall size displays. Furthermore, if other people would like to have a look at their own space while the main viewing is happening, other options such as opening a semi-transparent picture-in-picture in the corner instead of resizing the current screen being watched can be made available. In addition, deciding the relative sizes of family members' screens and their priority can be useful and requires an understanding of family dynamics before protocols are established. These interesting issues are left for further studies. Our main

results from the survey indicate that people feel that personal spaces on large displays are more useful for games, and users want their space sizes to stay static.

4.1.10 Discussion of User Study Results for Screen Real-estate Management

The main question that motivates this test is whether smart home technology can be useful to aid in better management of multiple personal spaces on a shared large display in the home. Although Tsandilas and Balakrishnan [36] found that users prefer to partition their own spaces in a Single Display Groupware situation, people in their homes may have different controllers, skills and requirements when making use of assistance using smart technology. For example, it takes considerable effort to use a remote control device to manipulate the positions of a personal space on a large display. As well, the use of a keyboard and mouse does not fit into the current practice of a typical home entertainment setting where people are coming and going as well as mixing passive and active interaction with the content. In this sense, smart placement of spaces according to users' presence and seating positions is one desirable solution to this issue as suggested by our study results. However, we also see that user performance and ratings were not significantly worse in the order-based condition. This suggests that while identification may be what cameras are useful for, placement may not be that significant. One major reason might be the width of the SMARTBoard used in the experiments is not large enough to make a difference. Also, the task is likely to be simple enough for users to cross sightlines without adding

considerable mental load, despite the discomfort of doing so. Alternatively, the random mistakes of the tracking algorithm (on average fewer than three times for four sessions due to temporal smoothing on the application side) may attribute to the lack of significance.

Like placement control, the manual sign-in/out process using the remote control devices could be troublesome, plus users constantly forget to sign out as was observed in our study. This is not so much a problem for the privacy of the user who leaves the room (although privacy could be a concern for certain tasks such as checking emails), but that their spaces are distracting and taking up the screen real-estate of the people still using the display. The automatic sign-out functionality is promising in solving this problem. Yet from the results, we can see that people perceive this approach does not provide a better sense of control and ownership of the space. This reflects that a person may not want their space to be signed out automatically, for example, just because he is away to grab a drink for two minutes. Also, the automatic showing of the space should the user enter the room is not always desired. Thus, there are a number of pragmatic issues that will need to be addressed before automatic sign-in/out can be integrated effectively.

As one cause of not getting significantly different results, the fidelity of the vision system complicates the usefulness of the smart environment. In our test, when the tracking system lost track of or misidentified a person, it reduced the users trust in the system quite dramatically and thus influenced their expectations and experience. For example, when errors occurred in the automatic sign-in/out condition, subjects subsequently tended to double-check if their spaces were removed as they stepped out. Therefore, we believe

that as an infrastructure for the human-display interface, the vision system needs to achieve as high a fidelity as possible with particular attention to achieving as low a false positive rate as possible. Nonetheless, some error is tolerable in the right context.

In summary, we think that automatic behavior for tracking-based placement and sign-in/out functionalities using a vision system hold some promise for interaction with a large display in the home, although, some pragmatic issues will be critical to address. Coupled with fidelity issues, the major lesson learned from the study is that the optimal solution to large display screen real-estate management may lie in a combined type of user control - the integration of the automatic system with the manual control, and that we should find out ways to strike a proper balance between the two. Thus, continued research into better ways to make manual control of the space, such as using gesture or different remote control mechanisms, is important as continued improvement of tracking technologies.

4.2 Study Two: Device Management for Large Display Interaction

As discussed in Chapter 1, we want to find new types of controllers that are cost-effective, and support multi-user ad-hoc use. Free-hand gestures can be an option enabled by the tracking infrastructure. However, as hand gesturing is lower in accuracy than key-press based remote controllers, we need to explore possible contexts where relatively low fidelity hand gestures will be preferred. The context we are going to test is in a group cooper-

ation task where a hybrid of high-fidelity and low-fidelity devices are used as complements to each other. We draw our idea from Guiad's Bimanual Theory [16], which found that in many everyday tasks involving two hands, such as hammering a nail into the wall, the two hands adopt different roles and perform asymmetric functions. The Dominant Hand (DH) performs the finer operations while the Non-Dominant Hand (NDH) provides rough guide for the Dominant Hand. We use the analogy of this theory in the group collaborative task, considering different group members using interactive devices of different fidelities may cooperate better than if everyone has the device with the same fidelity.

In an interactive environment, key-press based remote controllers are precise and reliable in issuing complex commands but are complicated to use, while hand movements can be mapped to express rough directional intentions rather than precise, sophisticated commands, such as pointing to a certain direction, or moving on-screen objects around on a large display. Therefore, we want to study if low-fidelity gesture controller can be a useful complement to high-fidelity remote controllers.

The scenario for this study is that three users sit in front of the large TV display, and cooperate in finding out differences between a pair of pictures. They will use high-fidelity remote controllers and low-fidelity gestural controllers to complete the task. Figure 4.9 illustrates this scenario. We will compare the experience between a condition where a group of people use all high-fidelity controllers versus a condition where a group of people use one high-fidelity controller and two other low-fidelity gesture controllers.

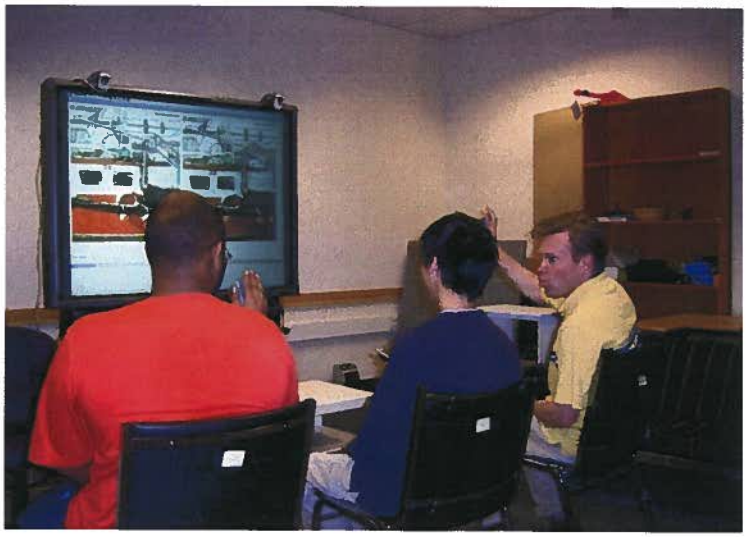

Figure 4.9: Device Management Study Setup.

4.2.1 Hypothesis

The hypothesis of this study is:

Users cooperate better in the group of hybrid devices (a master control device and several low-fidelity gestural devices) than homogeneous high-fidelity devices.

4.2.2 Independent Variables

Assuming groups of three subjects take part in our study. The independent variable is the configuration of control devices in the group. There are two conditions for this variable, described as below.

Condition 1. (Homogeneous Control) Everyone holds a mobile phone that can perform both coarse-grain and fine-grain control.

Condition 2. (Hybrid Control) Only one person can have the mobile phone with only fine-grain control while two others can wave color pads to the camera to perform coarse-grain control. This part makes use of the hand tracking infrastructure.

We use a within group design for the experiment.

4.2.3 Apparatus

The setup of the large display and seating in this study is the same as in Study One. We made two types of controllers: The high-fidelity controller, a mobile phone, whose keys are programmed so that users can perform fine-grain control as well as the coarse-grain directional control, and the hand-worn color pad, which serves as the low-fidelity coarse-grain control.

Figure 4.10: High-fidelity Device for Device Management: The Mobile Phone.

As is shown in Figure 4.10, the mobile phone can perform the following two types of control:

1. *Fine-grain control* that involves moving the hand-shaped cursor around in the view with the 5-way directional pad, and confirming the decision with

the central key labelled OK.

2. *Coarse-grain control* that involves scrolling both the pictures upward, downward, leftward and rightward together in the picture window of limited size with the four keys on the numeric keypad labelled with arrows.

We separate the physical locations of the fine-grain control and coarse-grain control, so as to help the users mentally distinguish the two control types. Users holding the mobile phone have access to both types of control in the homogeneous condition. The coarse-grain control is disabled in the hybrid control condition, replaced by gesture control. The keys of the mobile phone labelled "Next" and "Back"are for advancing the pictures and going back to previous pictures when needed.

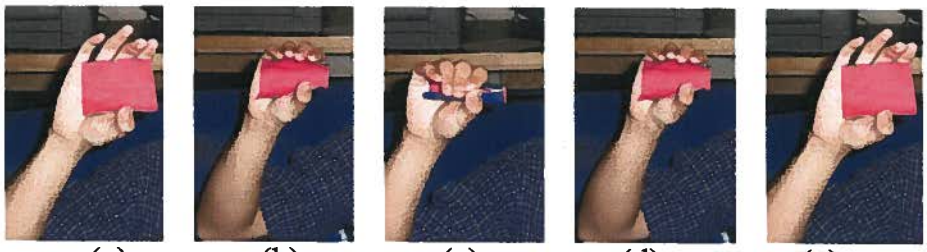

Figure 4.11: Low-fidelity Device for Device Management: The Color Pad. (a) - (e) illustrates a "clutch": from engaging in interaction to hiding the pad from the camera view, to engaging again. A "clutch" helps with users' transition from one movement to another. (a) Engage in interaction. (b) Prepare to hide the pad from view. (c) Hide the pad from view. (d) Prepare to engage in interaction. (e) Engage in interaction again.

Another "controller" is a uniformly colored pad for the users to hold in one hand to perform only *coarse-grain control* - gesturing to the camera to scroll the pictures, as is shown in Figure 4.11(a). Moving the hand upward

results in scrolling to the upper part of the picture, and accordingly with the downward, leftward and rightward movements. This controller is made of two layers. A clear plastic film serves as the base, and a colorful latex pad is attached on top of the base. This type of controller makes it easy for the users to hide the pad by closing their hand, and show the pad by opening their hand, simulating bare-hand gesturing. The opening and closing of the hand serves as a clutch (Figure 4.11) so that users can easily start and stop the commands. Each color pad is distinct in color in our experiment. While this is not how we expect an actual end-device to work, it mimics the functionality that will be required of a vision system that can do simple gesture recognition.

4.2.4 Application

The application is a game to spot differences in pairs of pictures shown in Figure 4.12. It shows a pair of photos which look similar but contain several differences. The goal is to observe the two photos and find out all the differences between them. There is one hand-shaped cursor on the left picture that can be moved to a place of difference by pressing directional keys on a mobile phone. Users can mark the difference by pressing the OK key on the mobile phone, and a red circle will appear at the cursor location. They can scroll the picture in the limited viewing window with the mobile phone or with the color pads. Both pictures are translated by the same amount when being scrolled. If multiple users try to move the cursor or the picture at the same time, the final movement will be the vector sum of all movements. There is a progress bar at the bottom of the picture pair to

indicate how much time is left. A timeout is set to 300 seconds per picture. If the group is unable to finish finding the differences within the timeout period, a timeout screen will appear and the group will need to advance to the next pair of pictures. There are also red bars on the four edges of the photos to prevent the users from further scrolling when they reach the limit of the image. Each group completes five pairs of pictures containing 24 differences in each condition. Each group needs to go through 2 conditions. In the hybrid control condition (Figure 4.12 Right), three color patches on the screen gives the visual feedback of whether the color pads are being detected by the computer.

Figure 4.12: Application for Device Management. Left: Homogeneous Condition. Right: Hybrid Condition.

4.2.5 Task: Difference Spotter

The task for the users are to cooperate with group members to find the differences in each pairs of pictures as quickly as possible, and make as few mistakes as possible.

This task is set up to provide different ways for people to cooperate and complete the task. We anticipate that in the homogeneous control condition,

the group will often be distracted in finding the difference as everyone has the control, while in the hybrid control condition, we hope to observe a higher level of cooperation, but possibly arm fatigue from people who control the color pad. We are looking to see whether users are more efficient in completing the task with hybrid control, whether they will cooperate better, and if they will have a better gaming experience.

4.2.6 Participants

Twenty-four subjects (16 male, 8 female) between the ages of 20 to 40, took part in this test. They are mostly UBC students and research fellows from majors ranging from Science and Engineering to Arts and Business. The subjects were divided into 8 groups of 3 subjects. During recruitment we encouraged subjects to bring their friends, so 7 groups out of 8 ended up composed of friends, lab-mates or house-mates, which made our test more realistic as we expect people in the home to know each other well. Twelve of them use computers over 40 hours per week, and 13 of them watch TV programs for less than 5 hours per week. Ten subjects had the experience of interacting with a large display TV once or twice, under circumstances such as playing video games in a friend's place, at a party, at a bar, in a meeting room, or participating in related research studies. Fourteen of them had never interacted with a large display before.

4.2.7 Procedure

The subjects played the game in two different sessions, each corresponding to one condition. The orders of conditions were counter-balanced across the

groups. In each session, the subjects were required to collaboratively find the differences in a total of 5 pairs of pictures.

Before each session, the subjects were provided a training session to familiarize them with the task and the input devices. The training task, as shown in Figure 4.13, has a pair of identical pictures with eight numbered squares and one central square. We asked each participant to take turns to use the mobile phone or the color pad to move the eight numbered squares into view. They could try multiple rounds until they thought they were skilled in controlling the device. They were also asked to use the mobile phone keypad to move the hand cursor around and mark the "differences". This procedure ensured that the novelty of the hand-controller would not be a main factor in the experiment.

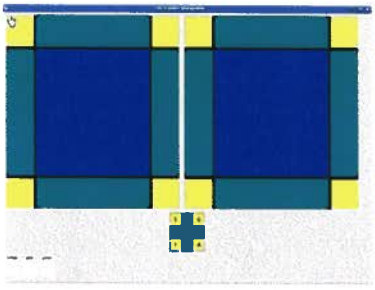

Figure 4.13: Training Task for Device Management.

4.2.8 Measures

We measured the average time spent on spotting differences for each pair of photos, and the correctness rates (number of correct differences out of a total of 24 differences) in each condition. Most importantly, we observed the type of cooperation exhibited in both groups. Subjects were also asked to

fill in questionnaires (Appendix A.2.4) after each session about their group collaboration. In addition, we interviewed them at the end of the test for their general comments.

4.2.9 Experimental Results for Device Management

The test results show that the average speed is comparable in both conditions, but the hybrid condition has a slightly higher correctness rate. Subjects rated the homogeneous control condition higher in efficiency, cooperation and fun, but lower in individual contribution. Ratings for "cooperation" in the homogeneous condition is significantly higher than in the hybrid condition. We observed an easier and more stable distributions of roles in the hybrid condition. Subjects consider the color pads hard to use, although they were able to retain their attention on the screen when using them. Detailed results are shown below.

The average speed and correctness rate of their performance is shown in Table 4.4. The hybrid condition has slightly (4.2%) higher correctness rate than the homogeneous condition, but the difference is not significant ($t(7) = -.002, p = 0.998$). The speed for both conditions is almost the same, again not significant ($t(7) = -1.302, p = 0.216$)

	Speed		Correctness Rate	
Device Combination Type	Mean(s)	SD	Mean	SD
Homogeneous Condition	193.76	52.07	86.55%	5.43%
Hybrid Condition	193.80	24.01	90.55%	4.55%

Table 4.4: Speed and Accuracy of Device Management Study.

Figure 4.14 shows that the subjects rated considerably higher in the

homogeneous control condition in efficiency, cooperation and fun. The observation and the post-test interview suggests that the reason for this is likely the stress from making the right moves in the hybrid condition. The identification of roles and consensus have the comparable ratings in both conditions. However, the homogeneous condition was rated lower in individual contribution.

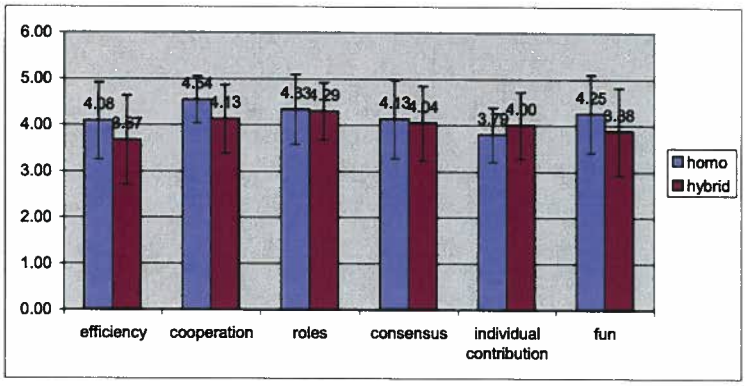

Figure 4.14: Comparison of User Ratings for Device Management Study.

In open-ended questions, all subjects who used the color pads as scrolling devices consider it harder to use than the mobile phone keypad, because it requires considerable physical stamina and skills to complete effective moves, despite the training given. The need to concentrate on the scrolling could prevent them from spotting the differences at the same time. However, one participant commented that he experienced a great deal of satisfaction and achievement when making gestural interaction because he had always been considered a physically awkward person.

In our observations, all groups were able to assign roles to people in the hybrid control condition, because the roles were pre-defined by the differ-

ences of devices. Specifically, the person with the high-fidelity controller marked the differences while the others moved the pictures around. In the homogeneous condition, 6 out of 8 groups were able to assign roles at the beginning of the game, therefore the cooperation was also quite smooth. They usually assigned roles in a common pattern: one subject pans the picture, one moves the cursor, a third spots the differences. The scrolling-subject moved when asked by a team member who spotted a difference outside the viewing window and wanted to go back. He initiated the move by asking other teammates' opinions first. The cursor-subject moved the cursor and pressed "OK" only when three of them confirmed the difference. The spotting-subject usually issued commands to the scrolling-subject and the cursor-subject. There were only two exceptions to this pattern. In one of the groups only one subject did all of the controlling in the homogeneous condition. In another group, one person was in charge of scrolling horizontally, and another, vertically. However, the common pattern afforded more flexibility in the homogeneous condition, since whoever spotted the difference would move the picture and the cursor without having to ask other people to do so. In this case, the different roles were not clearly demarcated. Whereas in the hybrid condition, subjects relied on each other in confirming the differences, and the roles were mostly static, because the barrier to changing role/devices was much higher. Only one group exchanged devices amongst each other because one subject did not feel comfortable making arm movements in the positions he was sitting.

For gesture control, it was observed that although subjects who used the color pad needed to pay special attention to make sure they make the right

movements, their attention on the screen was retained. People with remote controllers occasionally needed to look at the keypad when taking actions, especially if they had previously hit a wrong key.

Finally, in terms of the subjects' impressions of cooperation as elucidated from the questionnaire, statistical analysis shows that the difference of ratings on "cooperation" between homogeneous and hybrid conditions is statistically significant, $t(23) = 2.32, p = 0.03, Mean(homogeneous) = 4.54, SD(homogeneous) = 0.51, Mean(hybrid) = 4.13, SD(hybrid) = 0.74$. Other questionnaire differences were not found to be significant. Refer to Appendix A.3.2 for the questionnaire scores and user comments.

4.2.10 Discussion of User Study Results for Device Management

The test results suggest that hybrid devices could be a solution to the problem of a limited number of devices, and the complexity of their usage. However, their use is likely quite limited. If we attach a camera to the large display, bright-color pads can act as control devices for simple manipulations. Extending this concept suggests that when hand-gesture detection becomes reliable, it could be used for some limited types of control. The number of devices could scale to a fairly large number at low cost with this approach.

In the hybrid control condition, functionalities are pre-separated to different devices. As users rely on each other in completing the task, role-playing becomes a natural choice. The typical assignment of roles the subjects came up with is: one person focuses on moving the cursor and con-

firming the difference; at least one person scrolling the picture; and usually a third person concentrating on spotting the differences. From our observation, people adhere to their assigned roles (only one group of people wanted to exchange devices). Because of the clear separation of roles, the power and control of spotting the differences was very centralized, causing the subjects to make joint efforts in finding the difference. Different from what we expected, the one with the master controller did not dominate the decision making. Rather, the person assigned the role of spotting the differences directed the other two's activities. Whereas in the homogeneous condition, in spite of the flexibility of roles, the power and control was quite decentralized. Anyone who spotted, or thought there was a difference could direct the movement. Most often, because of social courtesy, other people would not object to the suggestion of scrolling the pictures. Therefore there were higher chances of going in the wrong direction. For example, when a person thinks he sees a difference in one area but is wrong, other members still tend to agree as they have no better choices at that moment. In this way, the homogeneous condition has a higher chance of making mistakes. This may be the case as the trend seems to head this direction. However, for this task and set of conditions, it appears that there is no significant difference.

Furthermore, the natural separation of roles in the hybrid condition increases the sense of participation and enhances the awareness in the group. We observed that all four groups that experienced the hybrid condition first adopted the role-assigning strategy in the homogeneous condition, while only two out of four groups that did the homogeneous condition first were able to assign roles. This means that the subjects realized that role-play, inherent in

the hybrid control condition, is natural and useful in this particular type of collaborative task. "Individual contribution" is rated higher in the hybrid condition, which is an indication of their better sense of individual value and participation in the collaboration. These results suggest that for some tasks, people will be able to negotiate protocols for using shared displays with different controllers that are of different fidelities. This will likely be important to the acceptance of shared displays in the home, as protocols for controlling the screen needs to be established, and it is likely that not every member of the family will always have a high-fidelity controller.

Apart from the fact that the hybrid devices increase individual contribution and may reduce error rates, there is always a trade-off between these merits and the intrinsic property of gesture interaction. Subjects rated efficiency in homogeneous condition as higher, although in reality efficiency in the hybrid condition is comparable. They also rated cooperation and fun higher in the homogeneous condition. From the post-test interview, we think that their perception of the above indices is largely compromised by the fatigue in controlling the pads in the hybrid condition. People felt that the color pads were difficult to control, although they thought that they could control well after the training. In more ad-hoc use during the process of the game, they still needed a great deal of effort to move to the right direction. The recognition system does complicate matters as it imposes constraints on the subject, specifically, sitting straight in the chair, not occluding each other, showing the pad to the camera, not waving out of the camera view, and so on. Therefore when the subjects reached their physical boundary, or were outside their own "active region" of gestures defined by our track-

ing program, they needed to hide the pad and start over. Some subjects were very exhausted by having to complete effective movements with all the constraints in mind. Ultimately, this difficulty stressed the subjects, consequently leaving a poor impression rather than a relaxed experience. We believe this will generally be true of any gesture-based system that requires regular free-hand movement to control the activity on the display.

This stressful experience could easily outweigh the feeling of novelty and excitement, and better cooperation. Some people simply commented "It didn't work well.", despite the program working at reasonable fidelity, only making mistakes at times. The experiment experience elucidates that the requirement of gestural interaction, coupled with technical imperfection of a vision system could still be a considerable barrier for user experience. The gesture interface, even with high accuracy, requires skill to make it more appropriate for gaming interfaces that require high levels of physical and mental effort. However, in regular screen control activities, only simple, once-in-awhile usage is recommended as accorded by our experiments. Keeping it simple allows people to easily remember and execute the gesture, and have the system recognize it. Making it once-in-awhile, such as channel changing, reduces the chance of fatigue setting in.

4.3 Summary

We presented two studies about screen real-estate management and device management in a multi-user dynamic interactive context. The results of these studies help us to determine whether a vision system like our tracking

system can be useful, and what it is useful for. The studies on identification and placement using either manual or automatic approaches show that for screen real-estate management, a method combining automatic and manual means can be useful. It will take advantage of automatic placement and identification facilitated by the tracking infrastructure, and also, the manual control enabled by the remote controller allow for a greater sense of control.

The device management results show that the even though hybrid control devices could be useful for centralizing control in a group task and reducing errors, it is only useful for simple and once-in-awhile interaction rather than usual control. More efforts are necessary to improve the fidelity of the hand-tracking infrastructure, and to further discover types of manipulation on the display that take better advantage of hand gesturing.

Chapter 5

Conclusion and Future Work

In this chapter, we summarize the work done in this thesis and the contributions made, including building up the real-time vision-based tracking system for an unconstrained interactive environment with reasonable fidelity, and carrying out extensive user studies to explore the usefulness of the vision system. We integrate lessons learned in terms of algorithm, system and user interface design into the discussion of future work, which concludes the thesis.

5.1 Summary of Thesis

The major contributions of the thesis are developing a multi-user identification and tracking system, and carrying out user studies to investigate the screen real-estate issues and device management issues in multi-user interaction with in-home large displays.

We began the thesis by introducing the research questions and goals. Due to their increasing size and functionalities, TVs are becoming the center of entertainment at home, and multiple family users tend to share the display for various other tasks besides simply watching TV programs and movies. They can use the TV display as an interactive space for both personal and

group entertainment activities.

For personal use, individual family members can have personal media contents stored on a server, and displayed in a virtual personal space on the TV for viewing and manipulation. This kind of display requires personal identity input and arrangement of space locations. In a more practical scenario, users come and go in the interactive area very often. It is difficult to use conventional remote controllers to perform signing-in/out functions or space placement according to personal preferences. A system able to sense the presence and locations of the interacting users may be able to assist the screen real-estate management for multiple users.

Further, conventional remote control devices seem to be less appropriate and more cumbersome to use for the growing number and complexity of functionalities. We need to explore novel types of controllers that facilitate more natural, ad-hoc control. Gesture interpretation systems would be a good choice, yet they are known for low fidelity and complicated setup process. Therefore, a group of control devices of complementary functionality might be preferred for group collaborative tasks such as gaming.

We proposed that a vision-based sensing and recognition system may be able to support an in-home large interactive display in that it can unite the functionalities of both human sensing and gesture sensing needed above. However, due to the challenges in a vision system, whether it holds promise in the future interactive TV system from the users' point of view is yet to be discovered, and is the focus of our thesis. Study results from the vision system can reveal the answers to usability issues in our current configuration, and provide design lessons for a generic vision system for future large display

TVs.

We pinpointed our research goals by discussing two motivating scenarios about multiple user interaction with the large display in dynamic scenes. We then put forward three hypotheses:

1. Users prefer to be automatically signed-in and signed-out of his space rather than manually.

2. Aligning a user's workspace on the screen according to his relative horizontal physical location with other users helps improve his task performance and overall interaction experience.

3. Users cooperate better in the group of hybrid devices (a master control device and several low-fidelity gestural devices) than homogeneous high- fidelity devices.

We then reviewed the literature regarding vision-based systems used in human computer interfaces, the interfaces between users and front panel screens in particular. We have found that most current vision systems for interaction are confined to single-user, high-resolution, close-range system, or multi-user system with the help of other devices than the sole vision system. Meanwhile, the CSCW community has a wide range of discussion of user experience of sharing and collaborating around the large screen display, such as proximity to the display, territoriality, and group collaboration on a single display, but a lot of these discussions are about tabletop displays. There is a lack of literature that explores the usefulness of a vision system in supporting a smart environment, in particular multiple users' sharing and collaborating on a large upright display.

To investigate the usefulness of a vision system, we developed a system (described in Chapter 3) that uses computer vision technology to identify and keep track of multiple users in real time in an indoor environment. The main algorithm of the system combines face recognition and color histogram tracking with Linear Programming as a temporal smoothing approach. Then we applied an improved Camshift approach to detect hand movement based on the multi-user tracking program. The program was optimized to meet the interactive rates. A Python interface was built to handle the communication among the tracking infrastructure, other devices, and applications for following user studies. Test results show that this algorithmic infrastructure provides sufficient fidelity for an interface with large display.

In Chapter 4, we described two applications regarding large display screen real-estate management and device management for testing the hypotheses. User study results reveal that the vision system holds promise for automatically placing users' personal spaces on the display according to their seating positions, and also for automatically signing in and out users depending on whether they are in the room or not. However, users would also like to maintain certain level of control when they are not satisfied with what the smart system has done. Furthermore, we found that bare-hand gestural interaction can be used for simple, once-in-awhile interaction with the large display, and the intrinsic properties of gesture interaction, complicated by technical imperfections, can compromise the user experience a great deal. However, it might help in a group collaboration scenario where one person has the master, high-fidelity control while others interact with the gesture input, in centralizing the control and reducing errors.

5.2 Future Work

Although studies in this thesis revealed answers to some interesting questions, there are several issues to explore as an extension of this work.

It is of great importance to improve the fidelity of the computer vision algorithm to make it more robust in more natural settings. The current algorithm works well when users look and dress distinctly, but more sophisticated appearance representation needs to be established, and geometric models of humans may also help. Background models need to be established to account for lighting changes. A second camera (like one of the two cameras shown in Figure 1.1, although only one camera was used in this thesis) would be helpful to validate the tracking results, enlarge the view angle, and solve discontinuity in time of occlusions. However, calibration and synchronization problems needs to be further addressed.

Another field worth exploring is the hand detection and gesture recognition. More efficient and reliable algorithms are called for to locate hands instead of color pads, and temporal clues should also be used. At the same time, to distinguish who the hands belong to needs geometric models and constraints. Better mechanisms regarding gesture recognition needs to be used, especially to understand the start and stop of a gesture.

Once the user tracking and hand tracking program achieve higher fidelity, we need to re-explore the design space for vision-based interfaces, taking into account the user study results we already have. For screen real-estate management, we may need to come up with ways to provide users certain level of control for their space sizes, and resizing mechanism to use (to stay static,

or change dynamically). Priorities could be applied depending on their task type, the amount of time needed, as well as family and social dynamics. For device management, we need to further investigate whether users prefer bare-hand gestures or gesture with a device in hand when interacting with the large display. Also, what functions are best for gesture interaction, and what gestures to map to those functions are important topics.

In summary, this thesis is one of the initial works on exploring the usefulness of a vision-based system for interaction with large displays at home. It reveals the large challenging research space that requires joint efforts from researchers in applied vision and human computer interaction.

Bibliography

[1] Nintendo. http://www.nintendo.com (accessed April, 2008).

[2] Open source computer vision library (opencv). http://www.intel.com/technology/computing/opencv (accessed April, 2007).

[3] Panasonic consumer electronics show. http://www.panasonic.com/cesshow (accessed April, 2008).

[4] Vicon. http://www.vicon.com (accessed May, 2008).

[5] N. Arksey. Exploring the design space for concurrent use of personal and large displays for in-home collaboration. Master's thesis, The University of British Columbia, Canada, 2007.

[6] T. Baudel and M. Beaudouin-Lafon. Charade: remote control of objects using free-hand gestures. *Communications of the ACM*, 36(7):28–35, 1993.

[7] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *CVPR '06: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 744–750, 2006.

[8] K. Bernardin and R. Stiefelhagen. Audio-visual multi-person tracking and identification for smart environments. In *MULTIMEDIA '07: Proceedings of the 15th International Conference on Multimedia*, pages 661–670, 2007.

[9] J. P. Birnholtz, T. Grossman, C. Mak, and R. Balakrishnan. An exploratory study of input configuration and group process in a negotiation task using a large display. In *CHI '07: Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 91–100, 2007.

[10] G. R. Bradski. Real time face and object tracking as a component of a perceptual user interface. In *WACV '98: Proceedings of the 4th IEEE Workshop on Applications of Computer Vision*, page 214, 1998.

[11] M. Burton, R.Jenkins, P. Hancock, and D. White. Robust representations for face recognition: The power of averages. *Cognitive Psychology*, 51(3):256–284, 2005.

[12] K. R. Castleman. *Digital Image Processing.* Prentice Hall Press, Upper Saddle River, NJ, USA, second edition, 1996.

[13] D. Demirdjian and T. Darrell. 3-d articulated pose tracking for untethered diectic reference. In *ICMI '02: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 267, 2002.

[14] W. Freeman and C. Weissman. Television control by hand gestures. In *IEEE International Workshop on Automatic Face and Gesture Recognition*, 1995.

[15] D. Goddeau, E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, and V. Zue. Galaxy: A human-language interface to on-line travel information. In *Proceedings of the International Conference on Spoken Language Processing*, pages 707–710, 1994.

[16] Y. Guiard. Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a mode. *The Journal of Motor Behaviour*, 19(4):486–517, 1987.

[17] K. Hawkey, M. Kellar, D. Reilly, T. Whalen, and K. M. Inkpen. The proximity factor: impact of distance on co-located collaboration. In *GROUP '05: Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*, pages 31–40, 2005.

[18] E. Hosoya, H. Sato, M. Kitabata, I. Harada, H. Nojima, and A. Onozawa. Arm-pointer: 3d pointing interface for real-world interaction. In *ECCV Workshop on HCI*, pages 72–82, 2004.

[19] S. Izadi, H. Brignull, T. Rodden, Y. Rogers, and M. Underwood. Dynamo: a public interactive surface supporting the cooperative sharing and exchange of media. In *UIST '03: Proceedings of the 16th annual ACM symposium on User Interface Software and Technology*, pages 159–168, 2003.

[20] A. Jaimes and N. Sebe. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding.*, 108(1-2):116–134, 2007.

[21] H. Jiang, S. Fels, and J. Little. A linear programming approach for multiple object tracking. In *CVPR '07: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2007*, pages 1–8, 2007.

[22] M. Karam, J. Hare, P. Lewis, and M. C. Schraefel. Ambient gestures. Technical Report ECSTR-IAM06-001, Intelligence, Agents, Multimedia Group, University of Southampton, 2006.

[23] M. Karam and M. C. Schraefel. Investigating user tolerance for errors in vision-enabled gesture-based interactions. In *AVI '06: Proceedings of the working conference on Advanced visual interfaces*, pages 225–232, 2006.

[24] F. C. M. Kjeldsen. *Visual interpretation of hand gestures as a practical interface modality.* PhD thesis, Columbia University, New York, NY, USA, 1997.

[25] N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma. A real-time framework for natural multimodal interaction with large screen displays. In *ICMI '02: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 349, 2002.

[26] M.S. Lee, D. Weinshall, and E. Cohen-Solal. A computer vision system for on-screen item selection by finger pointing. In *CVPR '01: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1026–1033, 2001.

[27] H. Liu, L. Zhang, Z. Yu, H. Zha, and Y. Shi. Collaborative mean shift tracking based on multi-cue integration and auxiliary objects. In *ICIP 2007: Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 217–220, September 2007.

[28] M. R. Morris, A. Huang, A. Paepcke, and T. Winograd. Cooperative gestures: multi-user gestural interactions for co-located groupware. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 1201–1210, 2006.

[29] B. A. Myers, R. Bhatnagar, J. Nichols, C. H. Peck, D. Kong, R. Miller, and A. C. Long. Interacting at a distance: measuring the performance of laser pointers and other devices. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in Computing Systems*, pages 33–40, 2002.

[30] J. Oh and W. Stuerzlinger. Laser pointers as collaborative pointing devices. *Graphics Interface*, pages 141–149, 2002.

[31] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV '04: Proceedings of the European Conference on Computer Vision*, pages 28–39, 2004.

[32] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: a survey. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 239–248, 2006.

[33] A. M. Piper, E. O'Brien, M. R. Morris, and T. Winograd. Sides: a cooperative tabletop computer game for social skills development. In *CSCW '06: Proceedings of the 20th anniversary conference on Computer Supported Cooperative Work*, pages 1–10, 2006.

[34] Quixel Research. Advanced tv comparisons 2004. http://www.quixelresearch.com/main.php?menu=press, (accessed May, 2008).

[35] G. Shoemaker, A. Tang, and K. S. Booth. Shadow reaching: a new perspective on interaction for large displays. In *UIST '07: Proceedings of the 20th annual ACM symposium on User Interface Software and Technology*, pages 53–56, 2007.

[36] T. Tsandilas and R. Balakrishnan. An evaluation of techniques for reducing spatial interference in single display groupware. In *ECSCW'05: Proceedings of the ninth conference on European Conference on Computer Supported Cooperative Work*, pages 225–245, 2005.

[37] E. Tse, J. Histon, S. D. Scott, and S. Greenberg. Avoiding interference: how people use spatial separation and partitioning in sdg workspaces. In *CSCW '04: Proceedings of the ACM conference on Computer Supported Cooperative Work*, pages 252–261, 2004.

[38] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR'01: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.

[39] D. Vogel and R. Balakrishnan. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In *UIST '04: Proceedings of the 17th annual ACM symposium on User Interface Software and Technology*, pages 137–146, 2004.

[40] F. Vogt, J. Wong, B. A. Po, R. Argue, S. Sidney Fels, and K. S. Booth. Exploring collaboration with group pointer interaction. In *CGI '04: Proceedings of the Computer Graphics International*, pages 636–639, 2004.

[41] A. Williams, D. Ganesan, and A. Hanson. Aging in place: fall detection and localization in a distributed smart camera network. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 892–901, 2007.

[42] A. Wilson and N. Oliver. Gwindows: robust stereo vision for gesture-based control of windows. In *ICMI '03: Proceedings of the 5th international conference on Multimodal Interfaces*, pages 211–218, 2003.

[43] T. Zhang. memjpegdecoder. http://vast.uccs.edu/vast/zhangt.html, accessed March, 2008.

Appendix A

User Study Material

A.1 Material for the Screen Real-estate Management test

A.1.1 Example Answer Sheet

Answersheet 0-1-0

1. flower

   policeman

   runner

   building

   sea

2. flower

   sky

   bridge

   boat

   car

3. sun

   escalator

   boat

   people

   store

4. escalator

   girl

   easter egg

   car

   soap

5. girl

   toy

   sky

   shelf

   tree

6. toy

   baby

   nurse

   soldier

   picture

7. picture

   tree

   floor

   ballet dancer

   flower

8. fish

   dancer

   diver

   fish

   sea

9. table

   sea

   customer

   clock

   plant

10. plant

    car

    bike

    table

    tree

11. girl

    bike

    fish

    bottle

    road

12. tree

    bike

    grass

    lake

    apple

13. flower

    tree

    lake

    honey bee

    sky

14. person

    sea

    bowling ball

    floor

    bike

15. car

    building

    floor

    people

    boxing glove

16. tent

    skates

    people

    sun

    food

17. tree

    sky

    lawn

    umbrella

    bike

18. umbrella

    boat

    car

    sea

    sand

19. dog

    sky

    man

    camera

    bike

20. chair

    coke

    kids

    cards

    cat

A.2 Questionnaire

A.2.1 Demographic Questionnaire

Questionnaire

Part I: Demographic Questions

1. Which age group are you in?

   19 and under

   20 – 29

   30 – 39

   40 – 49

   50 – 59

   60 and above

2. Gender:

   Female

   Male

3. Occupation:

   Undergraduate student, please specify major:

   Graduate student, please specify major:

   Academic, please specify area:

   Technical in industry

   Management

Other, please specify:

4. How man hours per week do you spend on using a computer?

   Under 5

   10 – 20

   20 – 30

   30 – 40

   40 +

5. How many hours per week do you spend on watching TV?

   Under 5

   10 – 20

   20 – 30

   30 – 40

   40 +

6. How do you rate your experience of interacting with a large screen TV or display:

   Often, please specify in what circumstances:

   Only once or twice, please specify in what circumstances:

   Never

   Other, please specify:

A.2.2 Questionnaire for Identity Experiment

Manual sign-in/out condition

Questionnaire

Part II: User Experience Rating

Please rate the following statement according to your experience when interacting with the large screen TV:

SD – Strongly Disagree

D – Disagree

N – Neutral

A – Agree

SA – Strongly Agree

Questions for those who have completed the tasks that required them to log in their spaces by menu selection.

1. I was able to identify my space quickly and easily.

   SD D N A SA

2. I had enough control over my space.

   SD D N A SA

3. I felt my space belonged to me.

   SD D N A SA

4. I like using the menu selection to log in.

   SD D N A SA

5. I felt it cumbersome using the menu selection to log in.

   SD D N A SA

6. I didn't want to log out myself when I went out.

   SD D N A SA

7. When I went out and came back, it was hard to find my own space.

   SD D N A SA

8. When someone went out without logging himself out, I found it annoying.

   SD D N A SA

Part III: Open- ended Questions:

How do you like the way of using menu selection on a remote device to log in/out your space?

If you had logged someone else out when he was out of the study area, why did you do this? If you hadn't, did you ever want to? Why?

Automatic sign-in/out condition

Questionnaire

Part II: User Experience Rating

Please rate the following statement according to your experience when interacting with the large screen TV:

SD – Strongly Disagree

D – Disagree

N – Neutral

A – Agree

SA – Strongly Agree

Questions for those who have completed the tasks that required them to show their faces to the camera to log in their spaces:

1. I was able to identify my space quickly and easily.

   SD D N A SA

2. I had enough control over my space.

   SD D N A SA

3. I felt my space belonged to me.

   SD D N A SA

4. I felt easy to log in because I only needed to show my face to the camera.

   SD D N A SA

5. I felt it annoying to show my face to the camera to log in.

   SD D N A SA

6. I liked the functionality that my space is gone after I went out.

   SD D N A SA

7. When I went out and came back, it was hard to find my own space.

   SD D N A SA

8. I liked the fact that when someone went out, his space was gone by itself.

   SD D N A SA

Part III: Open- ended Questions:

How do you like the way of showing your face to log in the first time?

What do you think of the fact that our system automatically signed someone out when he was gone?

A.2.3 Questionnaire for Placement Experiment

Order-based condition

Questionnaire

Part II: User Experience Rating

Please rate the following statement according to your experience when interacting with the large screen TV:

SD – Strongly Disagree

D – Disagree

N – Neutral

A – Agree

SA – Strongly Agree

Questions for those who have completed the tasks in spaces arranged according to fixed policies.

1. I was able to identify my space quickly and easily.

   SD D N A SA

2. I found it easy to work in my own space.

   SD D N A SA

3. I was attentive to my work.

   SD D N A SA

4. I worked fast.

   SD D N A SA

5. I had enough control when I was performing the task.

   SD D N A SA

6. I felt my space belonged to me.

   SD D N A SA

7. When I went out and came back, it was hard to find my own space.

   SD D N A SA

8. I found it annoying when other users came in and went out causing my spaces to be relocated or resized.

   SD D N A SA

9. I found it hard to resume work when my space was resized or relocated.

   SD D N A SA

Part III: Open- ended Questions:

There are fixed policies of arranging the spaces. Were you able to figure out the policies?

What's your experience of having to cross the sight line with each other (not directly facing your own space)?

How do you like the interface?

Tracking-based condition

Questionnaire

Part II: User Experience Rating

Please rate the following statement according to your experience when interacting with the large screen TV:

SD – Strongly Disagree

D – Disagree

N – Neutral

A – Agree

SA – Strongly Agree

Questions for those who have completed the tasks in spaces changed according to the tracking infrastructure.

1. I was able to identify my space quickly and easily.

   SD D N A SA

2. I found it easy to work in my own space.

   SD D N A SA

3. I was attentive to my work.

   SD D N A SA

4. I worked fast.

   SD D N A SA

5. I had enough control when I was performing the task.

   SD D N A SA

6. I felt my space belonged to me.

   SD D N A SA

7. When I went out and came back, it was hard to find my own space.

   SD D N A SA

8. I found it annoying when other users came in and went out causing my spaces to be relocated or resized.

   SD D N A SA

9. I found it hard to resume work when my space was resized or relocated.

   SD D N A SA

Part III: Open- ended Questions:

Do you find this interface intuitive? Why?

Can you tolerate the mistakes and occasional flipping of space locations if there were any?

How do you like this interface?

General question:

Questionnaire

General Questions:

What do you think of the idea of having individual workspaces on an interactive TV?

A.2.4 Questionnaire for Device Management Study

Questionnaire

Part II: User Experience Rating

Please rate the following statement according to your experience when interacting with the large screen TV:

SD – Strongly Disagree

D – Disagree

N – Neutral

A – Agree

SA – Strongly Agree

Please rate the experience of performing tasks in the group you've just worked in.

Group 1:

1. I found we were efficient in completing the task.

   SD D N A SA

2. I found we cooperated well in the task

   SD D N A SA

3. The roles were established shortly after the task started.

   SD D N A SA

4. The consensus was easily reached.

   SD D N A SA

5. I felt to have contributed a lot.

   SD D N A SA

6. We had a lot of fun.

   SD D N A SA

Group 2:

1. I found we were efficient in completing the task.

   SD D N A SA

2. I found we cooperated well in the task.

   SD D N A SA

3. The roles were established shortly after the task started.

   SD D N A SA

4. The consensus was easily reached.

   SD D N A SA

5. I felt to have contributed a lot.

   SD D N A SA

6. We had a lot of fun.

   SD D N A SA

Part III: Open- ended Questions:

How do you feel about cooperating in both groups?

Were you able to adopt some strategies from the previous group you were tested in? If so, what were they? Did they turn out to be useful?

Did you feel like changing devices in the hybrid condition group (The group with one cell phone and color cards)?

A.3 User Feedback

A.3.1 Questionnaire Result for Screen Real-estate Management Study

Identity Experiment Results

Manual sign-in/out condition

Scores for rating questions

	SD	D	N	A	SA	Average
1. Identify	1	0	2	8	7	4.11
2. Control	0	2	2	9	5	3.94
3. Ownership	0	1	1	10	6	4.17
4. Like menu selection	0	9	3	4	2	2.94
5. Menu cumbersome	2	3	3	10	0	3.17
6. Don't want to log out	2	3	2	9	2	3.33
7. Hard to resume after coming back	2	9	6	1	0	2.33
8. Annoying when ppl don't log out	3	3	5	5	2	3.00

Table A.1: Scores for Manual Sign-in/out Condition in Identity Experiment

Answers to open-ended questions

How do you like the way of using menu selection on a remote device to log in/out your space?

It doesn't make any difference using it or not. I like the way my space shows on the TV automatically. I don't like to look at the menu selection on cell phone.

I find it much more convenient than having to look at the camera to register our faces to the computer.

It better assures myself of successful logging in/out.

It was annoying because you need to press the button twice.

Sometimes annoying because of the lagging.

It's not as good as automatic sign in using camera.

I find it annoying compared to automatic log in/out system. Probably because I was spoiled with the auto systems.

It's nice to know I can log other people out, but unless someone went out and totally forgot, I wouldn't care. Besides, I can see myself playing tricks on others.

Kind of inconvenient to click twice to log in/out.

It is unnecessary because it seems like its an extra step.

Good, more control.

I like it, however, I do find the fact that I can be logged out by someone else somewhat is annoying.

I prefer direct log-in to using menu.

It's easy but people may be able to sign in as me easily.

It was very simple but could live without it.

I like it.

Not very convenient. Have to keep remembering to do it.

If you had logged someone else out when he was out of the study area, why did you do this? If you hadn't, did you ever want to? Why?

No. I just concentrated on my working space. It doesn't distract me from my own space.

I would have logged someone out when he was out of the study area because by logging someone out, then I would be able to get a full-screen

view of the pictures on TV. I hadn't logged someone out because they always logged themselves out when they left the living room area.

I don't want to. Cuz I didn't even notice someone forgot to log out. I was so attentive to my own space, and it seemed not bother me a lot when someone forgot to sign out, so why I bother to do that for them.

I wanted to, but he remembered and came back quickly.

Because she should sign out and give other people more room. It's a way of respecting other people.

I didn't. Because seems they did not forget to log themselves out.

I did because I wanted to make use of the new feature. The process itself is a bit of annoying and time consuming.

No. I was focusing on my work, much not even realizing if someone forgot to log out. Maybe if I was losing at a game and is a sore loser, I'd do it.

I did log someone out because that person finished the test.

I didn't, but would want to since it made my screen bigger.

No I didn't have to. I want to, give myself more space.

I haven't but I would since it increases my screen size. Also, if in the event of a TV show it reduces the other show's distractions.

I want, I favor larger screen.

I wouldn't do that. Because I respect his/her private space.

I logged out the other person by accident. No, didn't want to. Was not paying attention to any other screen but my own.

No I never wanted to. I had enough space and had no need.

Didn't really notice who was logged out who wasn't.

Automatic sign-in/out condition

Scores for rating questions

	SD	D	N	A	SA	Average
1. Identify	0	0	5	7	5	4.00
2. Control	0	5	2	6	4	3.53
3. Ownership	0	4	3	5	5	3.65
4. Easy to log in	0	2	2	4	8	4.13
5. Annoying to log in by face	5	3	7	1	1	2.41
6. Like automatic sign out myself	0	0	1	13	3	4.12
7. Hard to resume after coming back	4	8	2	2	1	2.29
8. Like automatically sign out others	0	1	1	10	5	4.12

Table A.2: Scores for Automatic Sign-in/out Condition in Identity Experiment

Answers to open-ended questions

How do you like the way of showing your face to log in the first time?

Good. I don't have to worry about anything.

No so much It feels like my face is being scanned for security reasons.

It is fancy, interesting, but not very functional.

It made it easier instead of logging in each time.

It's very convenient that I didn't have to press any keys to log in.

Convenient.

Convenient that I don't have to press ok every time to log in or out.

I like that, no login or thinking required, just walk in, look, go.

It's great.

It was good. Because we didn't have to log in/out every time we left.

Cool. Doubts about how sensitive the system is.

Fairly neutral about it.

Not bad.

It's nice if it can recognize me.

Didn't mind.

Better. Nice that you dont have to keep doing it.

I think it's great. Painless and easy to understand what you need to do.

What do you think of the fact that our system automatically signed someone out when he was gone?

It's convenient that I dont have to concern about signing in and out.

It is more convenient than having to sign out first when someone was gone.

It is nice, because it is so normal for people to forget about signing out thus taking space especially in this kind of interface.

It was pretty cool and when that happened my picture would get bigger so I didn't have to pan left or right.

It's very efficient and saved a lot of trouble.

This is nice and I wish other people could be gone for longer time so that I don't need to move the picture to see the whole thing.

That's really great. Why press extra buttons if you don't have to?

This was good since it automatically increased my viewing angle whenever someone left the room, unlike the other cases when the user might forget to press ok before leaving.

Didn't notice, but would be great.

It's good, to provide more space to others.

It was good because it made sure that the people who were still in the

space could maximize their area.

Good and efficient.

Useful.

I like it. I can be more concentrated in my space.

It's great!

It seemed efficient and maybe a little spooky.

Very user friendly, Good for confidentiality as well.

Placement Experiment Results

Order-based condition

Scores for rating questions

	SD	D	N	A	SA	Average
1. Easy to identify	0	1	2	7	8	4.22
2. Easy to work	0	3	4	6	5	3.72
3. Attentive	0	0	1	11	6	4.28
4. Fast	0	0	7	8	3	3.78
5. Control	0	2	2	8	6	4.00
6. Ownership	0	2	4	8	4	3.78
7. Hard to resume after coming back	5	10	3	0	0	1.89
8. Annoying by coming and going	1	4	1	11	1	3.39
9. Hard to resume after resizing	2	7	0	9	0	2.89

Table A.3: Scores for Order-based Condition in Placement Experiment

Answers to open-ended questions

There are fixed policies of arranging the spaces. Were you able to figure out the policies?

Yes.

Yes.

No.

Yes.

Yes.

No.

Kinda.

Yes.

Not clear.

Kinda.

Kinda.

Yes.

Yes.

Yes.

I didn't pay attention to that.

No.

No.

No.

What's your experience of having to cross the sight line with each other (not directly facing your own space)?

Cannot see the pictures clearly. Need to concentrate much more than previous time.

I find it not convenient and I get tired after a while having to turn my face a little bit.

A little bit vague, too far.

It made it difficult to see because pictures were small.

It's a lot harder to concentrate and focus, need to pay a lot of attention.

Does not matter. My space is in the middle.

Didn't find it distracting, but having my screen farthest away from me made it difficult to see. The combination of distance and viewing angle meant I had to squint my eyes at times to find some of the objects.

Sometimes confusing.

Very inconvenient and a little irritating. I couldn't see clearly what is in my space until one or both other people have left.

It was ok, but would've been better if I was directly facing my own space. Actually I was facing my own space.

Not preferred, easier to look ahead.

Does not bother me.

No Big deal.

It's ok. Didn't pay attention to it.

Didn't even notice really.

At first I thought it would be annoying but it was fine.

Don't like it very much.

How do you like the interface?

Responds quickly, self controlling.

It's ok. I think its user friendly enough since there weren't too many buttons on the phone. Instructions have to be used to control the TV.

So so, too far from my own space.

It was manageable.

It's fun.

The cell phone interface was good. It was very responsive and intuitive to use.

When the image is shared, it is sometimes very hard to distinguish objects, (but) since the images were not videos, it was easy to focus on my own space.

It's good. After a few times it will be more interesting.

Very cool and fast responding. Nice resolution.

It was good.

Good although would be better (with more) space.

It's cheaper to have many TVs, and there is no privacy.

A little.

Neutral. I prefer to have a bigger space.

It seems to work well.

It's kind of nice to only have buttons. No need to worry about random glitches.

Didn't like it very much.

Tracking-based condition

Scores for rating questions

	SD	D	N	A	SA	Average
1. Easy to identify	0	0	0	12	6	4.33
2. Easy to work	0	1	1	10	6	4.17
3. Attentive	0	0	1	12	5	4.22
4. Fast	0	2	3	10	4	3.84
5. Control	0	2	1	9	6	4.06
6. Ownership	0	3	0	9	6	4.00
7. Hard to resume after coming back	9	6	2	1	0	1.72
8. Annoying by coming and going	2	3	2	7	4	3.44
9. Hard to resume after resizing	3	4	5	4	2	2.89

Table A.4: Scores for Tracking-based Condition in Placement Experiment

Answers to open-ended questions

Do you find this interface intuitive? Why?

Very exciting and interesting.

Yes, user-friendly.

Yes.

It's easy to adapt to, almost like when you are on the computer. Instead of having your mouse, you use a cell.

Maybe.

Yes. It changes upon request.

Yes if it works well without errors. Compared to the first trial, my workspace is much closer to me, making it easier for me to find and identify items on the screen.

I was allocated the middle square so not a problem.

Yes, easy to recognize the spaces.

Yes because it senses where I am and adjust the space to where I am located to help with viewing.

Yes, because our personal space corresponds to where we were sitting. Always in the same space/place.

It's intuitive in the sense that it seems natural to use the person's presence to start his/her show. On the other hand, people may not always want to sign into their space just because they are in front of the TV.

Not so much. The pictures look small.

Yes. The spaces are located in front of the user (according to the positions).

I suppose so. It was very easy and uncomplicated.

Yes it was easy to tell which picture was mine because my name was there and it was always on the left.

Sort of. It moves when you go.

Can you tolerate the mistakes and occasional flipping of space locations if there were any?

Yes.

Probably not if the mistakes occur too many times.

Yes, to some extent.

It was a bit disruptive to have to get up those times.

Yes.

Yes.

No. This was very distracting. When the program makes a mistake, thinking someone left the room when they didn't, it caused my screen to quickly increase and then decrease in size. This effect causes a bit of a

headache for me. Also, when it flips space locations, it takes time for the users to find their space again. This is irritating.

I just had to relocate it or wait until normal conditions to resume the task.

Yes.

It's much better than cross-sighting to the other area to look for my own space.

Yes.

Yes.

Within reasonable limit, yes.

I can.

One or two are fine.

Yes.

It's pretty annoying. I probably wouldn't be happy if this was happening while I was performing a task important to me.

Not really. It's very annoying.

How do you like this interface?

Controlled the task by myself, not depending on other people.

The screen adjustment might not be convenient if I were watching TV and not looking at pictures only.

Not hard to follow.

I like it, it was very interesting.

It's interesting to locate the objects in the pictures.

I hope this it not for the real TV.

As I said in the first question, it's good if there were no mistakes in identifying people.

Because of the random resizing and disappearance, I found it annoying.

It's better than the previous one.

Much better than the previous (non tracking) one. Although the pictures are not too big when all 3 people were logged in, it is much easier to spot the objects when it's closest to my location.

It was alright, but annoying when the size of the space kept changing.

Better than the last (non-tracking).

Intuitive, but need to be improved. There need to have more control over the screen's decisions.

Not bad, but I don't like the small screen.

It's better than the previous one, because the pictures are displayed right in front of me. But still, I do not like my space to be resized too often.

It was fun and easy.

I like it except for when the photos resize.

It's ok but I would never have this in my home.

A.3.2 Questionnaire Result for Device Management Study

Scores of rating questions for Homogeneous condition

	SD	D	N	A	SA	Average
1. Efficiency	0	1	4	11	8	4.08
2. Cooperation	0	0	0	11	13	4.54
3. Roles	0	1	1	11	11	4.33
4. Consensus	0	0	7	7	10	4.13
5. Individual contribution	0	0	7	15	2	3.79
6. Fun	0	1	3	9	11	4.25

Table A.5: Scores for Homogeneous Condition in Device Management Study

Scores of rating questions for Hybrid condition

	SD	D	N	A	SA	Average
1. Efficiency	0	4	4	12	4	3.67
2. Cooperation	0	0	5	11	8	4.13
3. Roles	0	0	2	13	9	4.29
4. Consensus	0	1	4	12	7	4.04
5. Individual contribution	0	0	6	12	6	4.00
6. Fun	1	0	6	11	6	3.88

Table A.6: Scores for Hybrid Condition in Device Management Study

Answers to open-ended questions

How do you feel about cooperating in both groups?

The cell phone keys are too tiny to be controlled properly. It took time to get used to moving the color cards with proper speed.

No optimal solution in (homo) group, since all three have the phone. However, we spoke rather than act. And communication takes time.

I feel very interesting in (hybrid) condition. Felt a little tired in (homo) condition although we organized efficiently.

We allowed one guy to control both panning and mouse (in homo).

I feel it easier to cooperate in the first (hybrid) group. In the (homo) group, although I had a cell phone, I think it was not very useful.

In homo group, all 3 focused on finding differences. Two can control cursor when spotted a difference. In hybrid group, one has to focus on panning. Have to tell the guys with the mobile. The homo group is easier to be controlled, and efficient if the task for each person is assigned properly.

The homo group is easier to operate. The second group is not running very smoothly due to some technical glitches. People in both groups can cooperate to finish the task, while the second group requires some more efforts.

In homo group, the cooperation is very well. In hybrid group, it is a little bit hard to keep same step.

The sensor result is not as stable as mobile device, it is hard to control.

I found we could easily learn how to cooperate efficiently after a short time of training. We learned to cooperate more efficiently in the hybrid group than in the homo group.

The homo group was much easier because we were all working with the same devices. The hybrid group I became frustrated more quickly with the other members of the group as they tried to move the screen around.

I felt was easier in the hybrid group, however in the homo group, when I couldn't see something, others would click and move for me.

We were able to divide the tasks and be flexible in changing. Changing roles turned out to be useful.

We cooperated well.

In the hybrid group, it was difficult to adjust the positions.

The homo group is more efficient in terms of cooperation.

I think the homo group was more interesting. And also moving pictures is easier.

Homo group is better. If the sensor is better detecting the movement of the pad then hybrid group is more fun. It would also be better if there is no delay between the remote control and the display of the picture. I needed to cover the pad fully by other hand to avoid the sensor to be detected.

The homo one was better because it was easy to do the horizontal and vertical movement in the later case.

(We learned from the hybrid group) to decide each one's role.

After trying out the first (hybrid) task, it was much easier to collaborate in the second (homo) task.

Were you able to adopt some strategies from the previous group you were tested in? If so, what were they? Did they turn out to be useful?

We assigned different roles. We allocated different responsibilities to difference people, so no conflict happened in tasks.

We follow similar strategies of cooperating: Someone would give advice, another makes the moves, then all would search in the window.

Everybody makes sure to take a special role.

We used coordinate system (12 o'clock) for communicating directions. In hybrid group it was harder to agree on the direction of panning. Have to speak out where you want to pan.

Fixing the roles avoids the interference, and is useful.

We tried to scan the picture systematically, but it was much more difficult, and we often ended up jumping all over the place.

We had one key person controlling the cursor although 3 of us have the ability to do so. The similar way of dividing tasks (as in the hybrid group) creates less confusion and makes completion of the task more efficient.

From the top left to right bottom corner. One person plays the control, one plays the movement, one just points out the diff.

We reached consensus of systematic scanning.

We decided to let one person move the photo while the other two move the cursor. It makes the operation less confusing.

To establish roles for group members.

I was detecting the difference in both tests because the machine didn't detect my card (Experimenter: It was making mistakes at the beginning because of the lighting, then it could detect but she wouldnt want to do so as the roles were already assigned.)

Each of us had their own role within the team. Only one of the two people with a card used the card which saved time.

In the homo group we started to look for differences in an orderly way, starting from the top-left corner and going down. One person moved the cursor, and the other scrolled the picture. We were more efficient in the second (homo) task.

Did you feel like changing devices in the hybrid condition group (The group with one cell phone and color cards)?

I prefer cell phone.

Cell phone control was easier with no delay.

I prefer the cell phone to the color card.

Relatively hard to make perfect moves with color cards. I didn't mind changing devices.

I don't feel like trading devices.

I don't want to change device. I enjoyed the role of controlling the color cards.

Color card control is a lot of pain. Better with two pointing devices and one color card.

Felt like changing devices.

The intelligent system using sensor is not as good as mobile protocol. I want to change and use cell phone which can make sure the successful rate.

I don't like the color cards. More convenient and controllable using the cell phone.

I was the one with the phone which I know is easier, but at times I wanted to do (everything) it all by myself as I figured I could do it more efficiently on my own.

No, but I thought I would be better at using the color cards than the cell phone.

I would like to change devices when my teammate seems to have difficulty moving the cursor.

No it doesn't matter.

Not really. The color cards were difficult for all.

I would like to change devices.

I don't like (the cards). It's hard to move the picture.

I want to change devices because I get to try different controlling system.

Yes we did change and it helped.

Yes (I felt like changing devices.)

No, I think it worked well.

General comments in the interview

In hybrid group, I had more sense of involvement, and achievement from moving the picture around.

My roles were the same anyway.

I would like the hybrid group if it works better.

Would prefer gestures if it works 100%.

There are still problems to distinguish whos doing what.

In homo group, everybody can pan, detect. Whoever wants to do it, does it.

We coordinate well with same devices.

We should come up with techniques both simple and accurate.

Regarding the roles, it was a matter of whether letting the techniques decide your roles or you decide for yourself. If people are close enough, they will coordinate well with homo devices.

Same for me. But we were flexible in trading the devices and helping each other.

It was difficult to hold the card you need to hold it gently.

When two people move at the same time, it caused confusion in the first (hybrid) game.

It (the color card) wasn't very accurate. We are used to traditional ways such as cell phones and remote controls.

To use the cell phone even if it works 100%. Because it was tiring.

Would like to try if we had a better device and something sensitive and easier to hold on.

The homo condition was easier because we learned from the first condition about dividing the roles. The card was still hard to control (although the program worked well).

Observation Notes

Group 1:

Homo condition: S stood up to help.

They negotiated: "Who will get the control? Who moves? Who click?"

Hybrid condition: Roles were assigned very naturally. S got the phone, M moved, and G spotted the differences.

Group 2:

Hybrid condition: roles were assigned naturally: "You pan, you use the cell phone, you spot the difference, and help panning sometimes." They chatted more about panning, because X wasn't sure about the direction of panning.

Homo condition: After learning from previous condition, they negotiated "This time, I pan" etc. L took the leadership by controlling the cursor movement, and verbalizing: "Stop!" and the guy spotting the difference "up, up, down, down".

They focused more on communication over the picture contents.

Group 3:

Homo condition: They came up with a lot of strategies. And they assigned the roles at the beginning "You move, you spot, you move the cursor."

"You scan top-down, you scan bottom-up".

Hybrid condition: "You pan." Because Z is the best panner.

Group 4:

Homo group: K does most talking.

Hybrid group: This group of people move the picture in a relatively random fashion. K did most of spotting, and asking other people to move the picture.

Group 5:

Hybrid group: K panned, J used the cellphone. Then L panned most of the time. The cursor person: "Move the picture to the left."

Homo group:

"Do you guys want to control the cursor?"

"Move the picture down!" - forgot she can do this by herself.

Group 6:

Hybrid group: The cell phone guy: "right? Down!!" "Dont move!" "Where do you wanna go?"

Homo group: "let's start from the top-left" "You move, you spot, you move cursor."

Group 7:

Homo group: the group decided to let one person control everything, and all others spot the differences. At first they forgot they could move the picture around.

The group issue command to the controlling person, and the person will ask for all others' opinions.

Before the hybrid condition, they decided that one person pan, one use the cell phone, and one spot the differences.

Group 8:

Hybrid condition: J pans, and spot at times; E controlled the cursor; A spotted the difference, and helped panning at times.

Homo condition:

E: "Can we set up the roles first?"

"One moves the cursor, one does horizontal movement, one does vertical movement"

E takes charge most of the time.

But their focus was more decentralized: "Im gonna go down". "Im gonna go from the corner".

A.4 UBC Behavioral Research Ethics Board

Certificate of Approval

CERTIFICATE OF APPROVAL - AMENDMENT & RENEWAL

PRINCIPAL INVESTIGATOR:	DEPARTMENT:	UBC BREB NUMBER:
Kellogg S. Booth	UBC/Science/Computer Science	H03-80151

INSTITUTION(S) WHERE RESEARCH WILL BE CARRIED OUT:

Institution	Site
UBC	Vancouver (excludes UBC Hospital)

Other locations where the research will be conducted:
N/A

CO-INVESTIGATOR(S):
Joanna McGrenere
Regan Mandryk
Ying Zhang
Wei You
Martin Matthias Finke
Sheelagh Carpendale
Lyn Bartram
Mark Hancock
Mani Golparr Fard
Madhav Nepal
Colin Swindells
Petra Neumann
J. Karen Parker
Joel Lanir
Mike Blackstock
Barry A. Po
Rodger J. Lea
Melanie Tory
Rachel A. Pottinger
Sheryl AS Staub-French
Tamara Munzner
Anthony Tang
Garth Shoemaker
Lu Yu
David Sprague
Sherman Lai
Sidney S. Fels

SPONSORING AGENCIES:
Canada Foundation for Innovation
Natural Sciences and Engineering Research Council of Canada (NSERC) - "Collaborative Visualization and Interaction in Ubiquitous Computing Environments"

PROJECT TITLE:
ARTIFACT: Advanced Research, Techniques, and Informatics for Future Advantages in Construction Technology

CERTIFICATE EXPIRY DATE: June 6, 2009

AMENDMENT(S):	**RENEWAL AND AMENDMENT APPROVAL DATE:**
	June 6, 2008

Document Name	Version	Date

The application for continuing ethical review and the amendment(s) for the above-named project have been reviewed and the procedures were found to be acceptable on ethical grounds for research involving human subjects.

Approval is issued on behalf of the Behavioural Research Ethics Board

Dr. M. Judith Lynam, Chair
Dr. Ken Craig, Chair
Dr. Jim Rupert, Associate Chair
Dr. Laurie Ford, Associate Chair
Dr. Daniel Salhani, Associate Chair
Dr. Anita Ho, Associate Chair

A.5 Consent Form

February, 2008

Department of Computer Science
201-2366 Main Mall
Vancouver, B.C., Canada V6T 1Z4
Tel: 604.822.3131 Fax: 604.822.2684
Tel: (604) 822-9289 Fax: (604) 822-5485
www.cs.ubc.ca

ARTIFACT: Advanced Research, Techniques, and Informatics for Future Advantages in Construction Technology

Principal Investigator:

Dr. Kellogg S. Booth, Professor, Department of Computer Science

Email: ksbooth@cs.ubc.ca Tel: (604) 822-8193

Co-Investigators:

Dr. Sidney Fels, Associate Professor, Media and Graphics Interdisciplinary Centre

Email: ssfels@ece.ubc.ca Tel: (604) 822-5338

Dr. Roger Lea, Adjunct Professor, Media and Graphics Interdisciplinary Centre

Email: XXXXXXXXXX@XXXX.XXX Tel: (604) XXX-XXXX

Dr. Matthias Finke, Postdoctoral Fellow, Media and Graphics Interdisciplinary Centre

Email: XXXXXXXXXX@XXXXX.XXX Tel: (604) XXX-XXXX

Wei You, Graduate Student, Media and Graphics Interdisciplinary Centre

Email: XXXX@XXX.XXX.XX Tel: (604) XXX-XXXX

This study is intended to show how collaborative computing applications are used in practice. You will be asked to use a prototype computer application to accomplish a computing task. Your participation will help us assess the usability of this prototype application. Computer images such as schematic diagrams or rendered blueprints will be presented on a computer display and you will be asked to manipulate these images through the use of an input device. You will be exposed to network cameras for the computers to analyze your images in real time so as to control the on-screen contents. We will record your performance and analyze how the application is used.

Audio and video recordings may be made of your performance. You may also be asked to complete a questionnaire that will help us assess your experience with computer technology and your

impressions of the prototype application. By consenting to participate in this study, you also consent to participation in the audio and video recordings, and the completion of the questionnaire form. Your total participation time will be no longer than about one hour. In exchange for your participation, you will be provided with access to research computing applications and tools to assist you while you are being observed.

We will ensure that all recorded data are accessible only by project investigators and are kept secure in a locked faculty office. All data from individual participants will be coded so that your anonymity will be protected in any publicly available reports, papers, and presentations that result from this work.

We intend for your experience in this study to be pleasant and stress-free. If you are uncomfortable, or are unhappy participating in this study, you are free to withdraw at any time, without any repercussions to you whatsoever. We would be pleased to explain to you the purpose and methods used in this study in academic detail after your participation has concluded, and to furnish you with our results when they become available.

This research study is funded by the Research Network program of the Natural Sciences and Engineering Research Council of Canada (NSERC) and/or Panasonic R&D Company of America. Portions of this research will be used for graduate theses. If you have any questions or desire further information with respect to this study, you may contact Dr. Kellogg S. Booth or one of his associates at (604) 822-8193. If you have any concerns about your rights or treatment in this or any other UBC experiment, you may contact the Research Subject Information Line in the UBC Office of Research Services at (604) 822-8598. You have been given a copy of this consent form for your records. You do not waive any legal rights by signing this consent form.

I consent to participate in this study under the above conditions:

Name (Please Print): _____

Signature: _____

Witness: _____

Date: _____