uCom:
spatial displays for visual awareness of remote locations

Ana Luisa de Araujo Santos

M.Sc. Electrical Engineering
Universidade Federal do Rio de Janeiro, March 2006
Eng. Electronics Engineering
Universidade Federal do Rio de Janeiro, March 2004

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, in partial
fulfillment of the requirements for the degree of
Master of Science at the
Massachusetts Institute of Technology
September 2009

Author **Ana Luisa de Araujo Santos**
Program in Media Arts and Sciences
August 28, 2009

Certified by **V. Michael Bove, Jr., Ph.D.**
Principal Research Scientist, MIT Media Laboratory
Thesis Supervisor

Accepted by **Deb Roy**
Chair, Departmental Committee on Graduate Studies
Program in Media Arts and Sciences

uCom:
spatial displays for visual awareness of remote locations

Ana Luisa de Araujo Santos

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, on August 28, 2009
in partial fulfillment of the requirements for the degree of
Master of Science at the Massachusetts Institute of Technology

Abstract

uCom enables remote users to be visually aware of each other using "spatial displays" - live views of a remote space assembled according to an estimate of the remote space's layout. The main elements of the system design are a 3D representation of each space and a multi-display physical setup. The 3D image-based representation of a space is composed of an aggregate of live video feeds acquired from multiple viewpoints and rendered in a graphical visualization resembling a 3D collage. Its navigation controls allow users to transition among the remote views, while maintaining a sense of how the images relate in 3D space. Additionally, the system uses a configurable set of displays to portray always-on visual connections with a remote site integrated into the local physical environment. The evaluation investigates to what extent the system improves users' understanding of the layout of a remote space.

Certified by **V. Michael Bove, Jr., Ph.D.**
Principal Research Scientist, MIT Media Laboratory
Thesis Supervisor

uCom:
spatial displays for visual awareness of remote locations

Ana Luisa de Araujo Santos

Thesis Reader **Joseph Paradiso, Ph.D.**
Associate Professor, The Media Laboratory
Massachusetts Institute of Technology

uCom:
spatial displays for visual awareness of remote locations

Ana Luisa de Araujo Santos

Thesis Reader **Ramesh Raskar, Ph.D.**
Associate Professor, The Media Laboratory
Massachusetts Institute of Technology

Acknowledgements

Mike Bove, my advisor, for accepting me into the Object-based Media Group, and giving me this great opportunity of spending two years at the MIT Media Lab. This whole experience has made me grow in so many different ways. I really appreciate it.

Joe Paradiso for his kindness, patience and useful comments to this thesis.

Ramesh Raskar for also being a reader, and giving me very useful advice on what research direction was worth pursuing. I also appreciate the inspiration from his previous work: "The Office of the Future".

A very special thanks to three particular friends who have directly contributed to this work: **Jim Barabas**, for all the mentoring on theory, coding and debugging, but most specially for taking his time to help me even when he was super busy himself; **Yannick Assogba**, for supporting my work but mostly for keeping me sane, being by my side at all moments with his calm and wise personality; **Quinn Smithwick** for all his kindness to share his immense knowledge, and for revising this text so carefully.

My great thanks to MIT Mental Health staff: **Dr. Haleh Rokni** and **Dr. Xiaolu Hsi**, my "thesis coach", for taking their time to listen to me and for giving me such wise advices. They were essential on keeping me sane over the last not-so-easy months of the thesis.

A special thanks to **Linda Peterson** and **Aaron Solle**. I appreciate their kindness and support, particularly on wrapping up this work. I also appreciate the help of **Gigi Shafer** while I was sorting things through to come to MIT.

MIT Writing Center staff, **Susan Spilecki**, for carefully reading and reviewing most of this thesis' text.

Kristin Hall for working her magic and so kindly taking care of our group.

The Media Lab sponsors, without whose resources this thesis could not exist.

The remaining OBMG fellows: **Brandon Taylor**, **Santiago Alfaro**, **Tim Wegner** and **Daniel Smalley**, for all the fun and particularly for their patience with the only girl in the group.

My friends and colleagues at the Media Lab, a community of incredible people. In particular, thanks to **Alex Dragulescu**, **Kate Hollenbach**, **Ankur Mani** and **Sajid Sadi** for sharing so much fun and a few tears. Another thanks also goes to **Ankit Mohan**, for sharing equipment and being so willing to help.

Thanks to some non-human resources that have helped me so much throughout this process: **Hayden library**, **back massage**, **Sidney-Pacific's study room**, **the Media Lab coffee machine**, **Jim Barabas' couch**, **Pandora's music**, **the sunlight in Cambridge**, and **my bicycle**.

Tadeu Ferreira, for keeping our friendship as close as it has always been. Thanks for being always there for me.

Miguel Quartin, my beloved boyfriend, for always telling me what I needed to hear, not exactly what I wanted to hear. He knows me better than anyone else.

My mom, **Ilka**, and dad, **Sergio**, for having always supported my decisions and for withstanding the physical distance over these past two years.

My godmother, **Cleia**, who I wish were here to celebrate this achievement.

Contents

List of Figures

1

Introduction

1.1 Motivation

The deployment of communication technologies in the last decades has enabled people present in different locations to communicate on a regular basis. Physically separated individuals can connect using multiple interfaces, based on text, video, audio or symbolic representations. The development and growth of the Internet have enabled a major breakthrough for those communication possibilities.

The growth of a globalized workforce and the communication technologies are mutually reinforcing. We have witnessed Nicholas Negroponte's [36] predictions come true: our lives definitely depend less upon being in a specific place at a specific time. Therefore we perceive both needs and opportunities in this scenario. The needs derive from the major spread of outsourced job opportunities throughout the world, which significantly increases communication among distant coworkers, friends and loved ones. The opportunity lies in making the most of the current technologies to enhance real-time connectivity among people.

1.2 Proposed Work

This project is named uCom, which stands for "ubiquitous Cameras, observant monitors." uCom envisions to enhance the feeling of being present in two locations simultaneously. This vision is put to work by enabling remote users to be visually aware of each other using "spatial displays", live views of a remote space assembled according to an estimate of the remote space's layout. Local displays can portray live video views of the remote space acquired from multiple viewpoints. A remote view can have two possible display formats: (1) individual live video views or (2) multiple live views assembled in a graphical visualization that resembles a 3D collage.

The main element of the system design is the intermediate 3D representation of a remote space. It is composed of an aggregate of live video feeds from multiple viewpoints of the scene, which are rendered in a graphical visualization resembling a 3D collage. uCom does not focus on synthesizing a photo-realistic view of the real world from multiple viewpoints, but it enables users to browse this collection of live video views in a 3D spatial context that provides a sense of the scene's geometry. uCom takes advantage of the scene reconstruction algorithm used by the Photo Tourism project [42] [43] [44]. This algorithm, named Bundler [1], computes estimates of camera poses and a sparse 3D model of the scene. Given these estimates, we compute the positions to which the images should be rendered in the 3D image-based model of the remote scene. Additional navigation controls allow users to transition among the images, while maintaining a sense of how they match in the 3D space.

uCom's ideal setup is composed of two physically separate spaces, each one with multiple cameras and multiple displays. The system can utilize nearly any cameras and displays that support Internet video streaming, whether off-the-shelf consumer electronics or professional video equipment. Our system poses few constraints on how to position the devices. Cameras can have nearly any position, as long as neighboring images overlap considerably enough to enable a sufficient 3D scene collage computation.

As a multi-display environment, uCom also focuses on how users interact with each display. We address this issue from two perspectives: (1) how users directly select what is shown in each display and (2) how users physically position the displays. First, our interaction design targets simplicity. We provide an easy-to-use remote control interface that allows users to select a display and switch the remote view it shows. Second, we allow users to freely arrange the available displays. We assume that users will tend to position displays where they can attract appropriate attention without being disruptive. We argue that, by providing users with a 3D image-based model of the remote scene, we are empowering them to understand the remote space's layout. Therefore, users will tend to position the remote views where they can be relevant to their particular interests.

In summary, uCom tries to minimally interfere with how a space is laid out while assisting users in creating a mental model of a remote space. It should be clear that uCom is not a videoconference system, rather it focuses on enhancing visual awareness of remote spaces. Therefore, any direct communication between remote users require auxiliary systems. Chapter 4 provides more details on uCom's implementation.

1.3 Contributions

Specific elements of our system have been explored in related research areas, such as the following: "media spaces", awareness systems, videoconferencing, telepresence, Computer Supported Cooperative Work (CSCW), groupware, multi-display applications, and peripheral displays, among others. Even though uCom draws upon these other areas, it differs from previous work in specific characteristics, and sometimes in application.

Several other projects have tried to help users create a mental model of a physical space by portraying immersive views of an environment. However, uCom is novel in that it uses live video from different viewpoints arranged according to a 3D image-based model of a scene. Additionally, we focus on conveying always-on visual awareness of a remote site seamlessly integrated into the physical environment. Regarding hardware, uCom differs from most previous projects by not requiring particular specifications or positioning. Our system is **not** limited to recreating a "window" into another space, at a specific location, with particular equipment, limited to a certain number of users, at a given moment. Instead, uCom is flexible and highly configurable. It can use almost any cameras and screen-enabled equipment available in the physical environment. uCom also scales in richness as more equipment is brought into the space.

To evaluate uCom's possible benefits, we investigate to what extent the system improves users' understanding of the layout of a remote space. Our tests evaluate both the 3D image-based model and the features that allow mapping remote video views to local displays. It should be noted that our evaluations are performed in the environment of an office building. Chapter 5 gives more details on uCom's evaluation method and results.

1.4 Thesis Structure

The organization of this thesis is outlined as follows. Chapter 2 describes the vision behind uCom and the motivations behind our design choices. Chapter 3 contextualizes uCom with respect to related work and emphasizes its particular innovations. Chapter 4 focuses on system features and implementation details. Chapter 5 presents the evaluation method and analyzes its results. Finally, we explore some of the future possibilities and discuss conclusions of the project in Chapter 6.

2

Concept

In this chapter, we explain the motivations behind our design choices. We start by presenting the vision behind ucom's and the strategies we use to accomplish it. Then, we discuss our main design principles, which lay the groundwork for contextualizing uCom with regard to related work.

2.1 Design Principles

uCom enables remote users to be visually aware of each other using "spatial displays" - live views of a remote space assembled according to an estimate of the remote space's layout. The system uses video in its attempt to convey real-time awareness of what happens in a remote location. Yet, we want to move beyond establishing direct conversations between remote users. So we focus solely on visually portraying the spaces. We wish the reader to interpret it as a visual connection between spaces, not between specific users.

The main strategies to accomplish our vision are:

- *Computing a 3D image-based model of the remote space that enables users' understanding of its geometry and spatial arrangement.* The 3D image-based model is composed of an aggregate of live video feeds from multiple viewpoints of the scene, which are rendered in a graphical visualization resembling a 3D collage. While looking at one of the views, a user should still see how this image geometrically fits with others in the related scene. And, while switching between views, the system should render a transition path that is faithful to the physical spatial relation between images. Chapters 4 and 5 show Figures of the 3D image-based model for some particular spaces used as examples.

- *Seamlessly integrating different views of a remote space into the everyday physical environment.* Users should be able to use fairly any available video displays to observe live video from the remote space in two possible display formats: an individual view or multiple live views assembled in a 3D image-based model of the remote scene. Users can observe any available live views of the remote space on fairly any displays located at the periphery of their attention or as a central piece in their work environment. We allow users to freely arrange the available displays, as we assume that users will tend to position displays where they can attract appropriate attention without being disruptive. We argue that, by providing users with a 3D image-based model of the remote scene, we are empowering them to understand the remote site's spatial arrangement. Placement of displays will naturally vary according to the available equipment and the geometric restrictions posed by walls, doors, windows or furniture. Ultimately, users will choose to assign a specific remote view to a specific local display based on their particular motivations, e.g., like following on a specific situation, checking a remote colleague's availability, or even to establish a static geometric correspondence between the spaces that makes sense for users regularly present there.

The main design principles behind uCom can be summarized in the following assertions:

- *uCom is about awareness, not communication.*

- *uCom assists users in creating a mental model of a remote space.*
- *uCom is installed in bounded spaces, initially focusing on two remote indoor workspaces.*
- *uCom is flexible and configurable.*
- *uCom exclusively uses computer assisted image-based techniques to model a space.*

2.1.1 uCom is about awareness, not communication.

We envision uCom as a system that enables users to be aware of what happens in a remote space, ultimately improving the feeling of being present in a remote location through visual cues. It is important to clarify that uCom does not intend to be a videoconferencing system. It rather focuses on enhancing visual awareness between the spaces. Therefore, any direct communication between remote users require auxiliary systems. Yet, we don't exclude the possibility of adding features to enable communication through uCom in our future implementations. But it's certainly not the current focus of the project.

2.1.2 uCom assists users in creating a mental model of a remote space.

uCom conveys a geometric representation of a remote space by: (1) creating an image-based model of the remote space; (2) embedding remote views into the local environment.

It portrays multiple viewpoints of live video from a remote space in a 3D interface that is coherent with the remote space's geometry. Users can browse and navigate through this 3D representation of the remote scene. It also allows users to decide where to appropriately display any of the views from a remote space, which can be shown on computer screens or televisions around their own physical space.

These piece of equipment can be either opportunistically available or dedicated to uCom. The goal is to create an immersive interface to a remote space. The use of multiple screens increase the displayable surface area, increasing the chance of creating spatially immersive displays.

It is noteworthy that uCom focuses on the 3D positioning of images to emphasize their spatial relationship, whether users are arranging computer screens in the local physical space or navigating the image-based 3D model of a remote space.

2.1.3 uCom is installed in bounded spaces, initially focusing on two remote indoor workspaces.

We define a "uCom space" or "uCom room" as a bounded architectural area with cameras, displays or both, controlled by computers able to connect to another remote uCom room through the Internet.

The ideal setup is composed of two uCom rooms, as this configuration conveys mutual awareness, prevents asymmetry of information and minimizes privacy concerns. If both spaces have multiple cameras and multiple displays, by seeing an individual in a remote space can imply that one can also be seen by others. After getting accustomed to uCom, users can relate to the familiar feeling of being in a public space.

It is worth mentioning that the current scope of this project is to install uCom spaces in workspaces, whether offices, meeting rooms or common work areas. But we can foresee future implementations of uCom in other types of bounded spaces: indoor and outdoor, private homes and public areas.

2.1.4 uCom is flexible and configurable.

Flexibility is one of the system's main characteristics. Equipment specification and position can change, creating a correspondence between two uCom rooms that best suit users' needs.

Regarding equipment specification, each uCom room should ideally have both cameras and displays controlled by computers connected to the Internet.. The system does not require a complex equipment setup. It can use off-the-shelf video cameras or webcams, and any displays, whether computer monitors or televisions. It can even use cell phones, if they support receiving or transmitting video streams. More sophisticated equipment can be incorporated if available, but we focus on taking advantage of pervasive off-the-shelf consumer electronics.

Cameras and displays don't have a pre-established location, but neighboring images should overlap to enable the computation of the room's 3D image-based model. We will further detail some "rules of thumb" for a minimum overlap between neighbor images, and features in Chapter 4. Cameras do not have permanently fixed positions. They can be modified by the users to meet their changing needs. But it should be noted that every time a camera is moved the system needs to recompute the 3D image-based model, which takes a few minutes. So, cameras should remain still most of the time. The displays, on the other hand, can be placed anywhere that fits users' needs. They can be central or peripheral to users' attention, like ambient displays on walls, or other computer monitors not being used at the moment. It should be noted that the more images, the higher the likelihood of a satisfactory 3D model. In addition, the better to enable a spatial understanding of the room.

The geometry of the two uCom spaces don't need to correspond or have similar scale or shape. It's at the user's discretion how to "map" the two rooms with the available equipment and the geometric restrictions presented by the room walls, doors, windows or furniture. Cameras and displays are not necessarily mapped in a one-to-one correspondence. So, the number of cameras in one uCom space can differ from the number of displays in the remote uCom space. Users can define which remote camera views are mapped to each local display or if multiple displays should show the same camera view.

The use of uCom can have multiple purposes or motivations. It can be used either to follow on a specific situation happening at the remote location real-time, to be more mindful of remote colleagues' availability in the context of their current activities, or even to establish a static geometric correspondence between the spaces that make sense for users regularly present there. In this case a user can, for instance, look right from his regular seat position and see the same view onto the remote space, creating a window into the other space.

2.1.5 uCom exclusively uses computer assisted image-based techniques to model a space.

A major element in uCom's design is its 3D intermediate representation of a remote space. It is definitely the system element that requires most of our attention. Yet, in order to make uCom flexible and scalable, we have decided to make its space modeling computation possible even while using regular image capture devices, i.e., cameras that can be easily found in our everyday environments. Therefore, it should be clear that uCom does not rely on any specialized equipment, such as: laser scanners, GPS, wireless signal triangulation, among others. Our scene model computation exclusively uses computer assisted image-based techniques to determine the scene geometry, the location and the orientation of cameras. Chapter 4 describes in details on how uCom's current prototype is implemented.

3

Related Work

In this chapter, we contextualize the previous discussion on the design considerations of uCom, presenting its similarities and differences with respect to related work. We shed light on particular aspects of uCom, and on other work that has inspired its development.

3.1 Introduction

While developing uCom, we reviewed related areas with which uCom is strongly or loosely connected. Specific elements of our system have similarities with previous projects and research areas which have been extensively explored, such as "media spaces", awareness systems, videoconferencing, telepresence, Computer Supported Cooperative Work (CSCW), groupware, multi-display applications, and peripheral displays, among others. Even though uCom draws upon these other areas, it differs from previous work both in specific characteristics, and sometimes in application.

To categorize the related work, this chapter covers four main areas: (1) spatial coherence between remote rooms, (2) immersive image-based models of real-world scenes, (3) multi-display applications, and (4) awareness applications. We describe the most relevant projects as follows.

3.2 Spatial Coherence Between Remote Rooms

Even though uCom is not a videoconference system, it explores a topic that is very frequently addressed in videoconferencing: how to convey a spatial connection between remote locations, ultimately enabling remote users to feel like sharing the same room.

There are multiple attempts to provide realism to videoconferencing, i.e., to create videoconference systems in which the remote room seems like an extension of the local room. Several systems have implemented immersion features, trying to replicate the stimuli users would naturally have if meeting face to face, or feeling like being in that other location. Life-sized images, shared meeting tables, eye contact, and spatial sound are common examples of "stimuli" telepresence systems try to provide. Several projects have implemented some, or all, of those features.

Several previous videoconferencing systems have tried to tackle this problem. Some projects create a virtual meeting scene and render all remote users' images in this synthetic environment. The idea is to enable users to feel immersed and share a common environment, even in a virtual world. Other projects approach the problem in a different way, focusing on establishing visual coherence between remote rooms. They connect rooms that have identical layout, furniture, lighting, etc, and align their images to the furniture. Some relevant examples are presented next.

3.2.1 Synthetic environments

Kauff et al [29] segment users' images from the real world background, and render them altogether in a synthetically generated 3D meeting scene. The project uses a head-tracking device to acquire the viewer's position, and renders the perspective view of the scene accordingly. Another project, Coliseum [11], implements a multiuser immersive videoconference in which five cameras attached to each computer acquire videostreams that are used to produce arbitrary perspective renderings. All participants are rendered in a shared synthetic environment, which aims to provide the participants with a sense of interacting in this common virtual world. Mulligan et al [34] render a 3D model of the world using binocular and trinocular stereo to create a view-independent scene model. Reflection of presence [9] is a videoconference application which focuses on collaboration around a visual artifact. Live video views of multiple participants are segmented from its original backgrounds, and then dynamically combined in a single synthetic environment. The scene background can show different media objects, like documents, images, prerecorded or live video through which users can collaboratively navigate in real time. Users' live views are dynamically combined into a single frame using varying transparency, position or scale that emphasize or de-emphasize the presence of each, i.e., each one's participation level in the interaction. Teleports [24], on the other hand, uses a full wall display surface to merge the physical space to a virtual synthetic environments for its videoconferencing application. The "display rooms" have one wall that is a "view port" into a virtual extension. The geometry and lighting of the physical space and its virtual synthetic extension are designed to closely match. Video images of remote participants are rendered into a virtual synthetic model, while the viewing position of the local participants is tracked, allowing imagery appearing on the wall display to be rendered from each participant's perspective.

3.2.2 Identical rooms

Other systems recreate visual coherence by connecting two almost identical meeting rooms, usually by simulating a shared meeting table. Cisco Telepresence^{TM} [3], is a current commercial videoconference product which aligns a meeting table in one room to the image of the remote room's table on life-sized screens. The system uses the same type of furniture, color of the walls, and lighting in both rooms. Multiple large screens, spatial audio, high communication bandwidth, and reliable technical support have stimulated the adoption of such systems, even at considerable cost. HP offers a similar commercial product: HP Halo^{TM} [5].

MultiView [37] is an example of a videoconference research project that simulates a continuous room in which two remote groups of people seem to be seating on opposite sides of a table. The system renders perspective correct views of the remote scene tailored to the position of each participant. The setup is comprised of cameras and projectors at multiple viewpoints, and a directional screen that controls who sees which image. They run experiments to research the effects of preserving eye contact, or gaze. The MAJIC [39] project, on the other hand, implements a multi-party videoconferencing system that projects life-sized videos of participants onto a large curved screen. It aims to convey the effect of users from various locations attending a meeting together around the same table. It also preserves eye contact and delivers directional sound.

3.2.3 Perspective correct views

It should be noted that several videoconferencing projects try to preserve eye contact between remote users, usually by rendering views of remote participants with perspective correct from each user's viewpoint. They tackle a common demand from users: feeling like they are looking in each others' eyes while using a videoconference system. Several of those systems limit one user per physical location, but others are capable of displaying perspectively correct remote views for multiple users present in one same physical location. Several videoconference systems try to mimic eye contact. A few relevant examples are MAJIC [39], Coliseum [11], MultiView [37], Teleports [24], [29] and [34].

3.2.4 How uCom addresses spatial coherence

A common aspect of the aforementioned examples is the attempt to create a very controlled geometric correspondence between remote spaces. Most approaches focus on pre-establishing the user's position, restricting the shape and size of remote rooms, and requiring a very constrained alignment between displays. uCom, on the other hand, addresses spatial coherence between remote spaces in a quite loose way.

First, we don't focus on creating spatial coherence *per se* between remote spaces. Our system is not limited to recreating a window into another office, positioned at a specific location, limited to a few number of participants at a given moment. We assist users in creating a mental model of a remote space through our 3D image-based model of the remote scene. Second, we enable users to easily display the views of the remote space on any equipment with available screens in their physical environment. The goal is to create an immersive interface to a remote space through these visual connections. Third, we allow uCom to scale in richness. As we try to make use of most cameras and displays available in workspaces, we don't want to impose many restrictions. Concerning the equipment used, as long as the cameras are positioned to appropriately compute the room's 3D image-based model, uCom users are allowed to reposition the displays at will. They can be central or peripheral to users' attention, like ambient displays on walls, or other computer monitors not being used at the moment.

We can state that uCom addresses spatial coherence between spaces as it allows 3D positioning of images to emphasize their spatial relationship. This happens whether while users are arranging computer screens in the local physical space, or while navigating the image-based 3D model of a remote space.

3.3 Immersive Image-based Models of Real-world Scenes

There has been extensive research on reconstructing real-world scenes. Several projects have tried to help users create a mental model of a physical space by portraying immersive views of an environment. For the scope of uCom, we are mostly interested in image-based models of real-world scenes created from multiple views acquired by diverse cameras.

We highlight related projects that use image-based models to convey immersion in a remote scene, as they are specifically relevant to uCom. Yet these projects differ in many aspects. First, the diverse kinds of equipments used. Some projects use regular off-the-shelf cameras, while others utilize specialized hardware, such as omnidirectional cameras with fish-eye lenses, head-mounted displays, among others. Second, the way multiple images are acquired. Some projects utilize images acquired through rotations around the camera's center of projection, a technique known as panorama or panorama stitching, which compose a single wide angle image. Other projects use multiple views of the same scene that have not necessarily been acquired from the same viewpoint. Their reconstruction method is usually referred as collage, photomosaic or image stitching. Third, they are used in multiple diverse applications, such as real-time telepresence systems, "3D movies", immersive 3D models, collages of still images, and so forth. Some of those projects are described as follows.

3.3.1 Using images acquired with the same center of projection

Some systems focus solely on creating panoramic videos. Teodosio et al. [51] aims to create navigable movies by using horizontal and vertical overlaps between frames. Their system acquires images using a video camera mounted on a computer-controlled pan-tilt mechanism, sampling the room by panning 360° from the center of the room. The result is a spherical projection of the movie. Other projects use live video streams with panoramic imaging to create telepresence systems, like the ones mentioned next. PanoCAST [49] creates a panoramic broadcasting system that combines panoramic video with the scene's 3D model. Video is captured by a spherically mounted camera and transmitted in real-time, while a remote user navigates the environment using a head-mounted display that maps parts of the spherical video onto a flat screen. Apple$^{\text{TM}}$QuickTime VR [1] (virtual reality) is a plugin to Apple$^{\text{TM}}$QuickTime that enables an immersive user experience using panoramic movies. It has two distinct panoramic modes: (1) cylindrical (a 360° image wrapped around the user's viewpoint), and (2) cubic (a cube of six 90° x 90° images surrounding the viewer). Users can navigate the 360° panorama with a mouse, tilting up and down, or selecting objects to be seen from different angles. Baldwin's [12] work acquires 360° of a remote location using a robot with a camera attached to a conic mirror. The system minimize the communication bandwidth by predicting the viewing direction of a robot's remote user. Ikeda et al. [27] create a telepresence system that uses a spherical immersive display. Navigation controls allow users to switch between view points in real-time by projecting different parts of the panoramic view.

Panoramas composed of still images, rather than video, are more frequent. For instance, Tomite et al. [52] enable a user to navigate a 3D model of a scene from an arbitrary viewpoint. 360° images are acquired by omnidirectional cameras placed at multiple positions. An omnidirectional image from any viewpoint is computed from three omnidirectional images acquired around the target position. Similar work is done by Fiala [19], but applied to a robot's navigation system exclusively using visual features. The robot recognizes an environment from a 3D scene structure pre-computed from multiple images captured by an omnidirectional camera. The 3D model is used to find other panoramic images taken in the vicinity, which are matched using SIFT keypoints [31] and computed using Structure from Motion (SfM).

Still panoramas are also used to create "virtual tours" throughout urban areas. Google Street View [4] [53] provides an interface to visualize street-level still images. It also enables navigation between images while preserving the context of the underlying street map. The current system version generates panoramic views from images acquired by vehicles equipped with 360° cameras and additional sen-

[1] Apple QuickTime VR: http://www.apple.com/quicktime/technologies/qtvr/

sors to compensate for the vehicle speed, acceleration and position on the scene computation. Google Street View is not the only project that has done it, but it's by far the one that has gained the highest popularity and largest coverage.

3.3.2 Using images from multiple viewpoints

A controlled method to acquire images from multiple viewpoints is by using arrays of cameras. For instance, Majumder et al. [32] implemented an immersive teleconference system using an array of cameras with approximately the same center of projection. The resulting image is rendered in real-time on a projection surface that surrounds the remote user. It resembles a panorama, but it is strictly a collage. Another example [38] uses a specialized camera array that can be physically flexed by the user to vary the composition of the scene. An automated method creates a collage from photos of a scene taken from different viewpoints with the camera array, conveying scene structure and camera motion. The images are aligned using least-squares formulation and similarity transforms by matching SIFT features [31].

On the other hand, scene collages do not necessarily require specialized hardware. A very relevant example is the Photo Tourism application, created by Snavely et al. [42] [43] [44], which innovates by allowing users to interactively browse and explore large unstructured collections of photographs. The main goal of the system is to evoke a sense of *presence* while browsing a collection of photographs from a scene. The underlying vision is to allow virtual tourism by taking advantage of the massive amounts of photos of touristic sites available on the Internet. The system can handle images captured from several viewpoints, levels of detail, resolution, lighting, seasons, decades, etc. The system works by automatically computing the viewpoint of each photograph and a sparse 3D model of the scene. A "Photo Explorer" interface renders the images and allows 3D navigation of the set of images according to the estimated 3D geometry of the related scene. Its main user interface navigation controls are: flying around the scene in 3D by blending neighbor images, browsing the scene by showing more images that contain a specific object or part of the scene, showing from where the photo was taken, and transferring object annotations from similar images. It can also uses geo-referenced absolute coordinates of the cameras to align - rotate, translate and scale - the estimated scene's model with digital elevation maps provided by external sources. Yet this process is not relevant to our work as we are only interested in relative positions of each camera, rather than absolute coordinates. Snavely et al.'s work has gone through several iterations on the user interface design [42] [43] [44]. Their work ultimately led to a sophisticated user interface provided by the very popular Microsoft Photosynth [7].

Another relevant project, "The Office of the Future, also aimed at using multiple views to connect remote spaces. Synchronized cameras would capture reflectance information of all visible surfaces in an office space. Real-time computer vision processing would dynamically extract per-pixel depth and reflectance information of all visible surfaces in the office, including walls, furniture, objects and people. Then, images would be projected directly onto the surfaces, and any objects could be used as a display surface. All office lights would be replaced by projectors, in order to precisely control the office lighting. The complexity of the system was mostly due to having time-variant characteristics, such as temperature and vibration, alignment, projector calibration, color and intensity balance, among others. The vision was to also be able to transmit the dynamic image models over a network for display at a remote location. Local and remote offices appear to be physically joined together along a common junction, such as a specific wall assigned as a spatially immersive displays. The Office of the Future's vision was later partially implemented with a reconstruction of a real but static remote office [54]. The implementation of the system involves the following sequence of steps: (1) acquisition of the remote scene's depth and color; (2) modeling of the scene into textured 3D geometry; (3) tracking of the user's eye positions with a head-mounted optical tracker; (4) rendering of the models based on the tracked user viewing position; (5) presentation in stereo for the user, who wears polarized glasses. The resulting 3D model of the remote office was rendered and projected onto a flat surface.

3.3.3 How uCom creates an immersive image-based models of a real-world scene

It is noteworthy that uCom focuses on rendering live video from a remote space, instead of photos. uCom allows that while looking at one video view, a user can still see how if fits with the others in 3D space. When the user transitions from one image to another, uCom renders smooth 3D transition paths between the images. Similarly to Photo Tourism, uCom does not focus on synthesizing a photorealistic view of the world from multiple viewpoints, but it enables users to browse a collection of videos or images in a 3D spatial context that provides a sense of the geometry of the referred scene. uCom makes use of the scene reconstruction algorithm provided by Photo Tourism's Bundler code [1], which computes estimates of the camera pose and sparse model of the scene. Although our rendering engine and navigation controls were custom built, they also draw upon Photo Tourism's idea. More details on how Photo Tourism and Bundler work are described in Chapter 4.

In uCom's case we are not interested in using special image capture hardware, unlike some of the aforementioned projects. Specifically the projects that compute panoramic views of a scene require a special setup or specific cameras capable of acquiring omnidirectional images. We want to create visualizations of a scene by using images acquired from multiple viewpoints, with any types of cameras available. In addition, uCom doesn't necessarily restrict camera's locations, but it requires neighboring images to overlap so as to enable the computation of the room's 3D image-based model. On the image display side, the hardware differs from most previous projects for not requiring specific positioning or specification. uCom also does not use a dedicated projection surface to render views of the remote space. It uses regular displays.

We would also like to highlight uCom's similarity with "The Office of the Future" [41]. Both projects make use of multiple camera views and focus on having the system installed in a regular workspace rather than creating a new dedicated space "down the hall", like a videoconference room.

3.4 Multi-display Applications

uCom draws inspiration from a series of multi-display applications, such as interactive meeting rooms and spatially immersive displays. A uCom room is undoubtedly a multi-display environment as it incorporates displays of multiple kinds and sizes, whether TVs, computer monitors, or mobile devices. Yet our system is very unconstrained regarding the use of devices, we allow users to decide how and where to position the displays, and which images each one should show. Yet we will describe some related projects in which the use of displays is more restricted requiring placement at specific positions. The similarity between these projects and uCom lies in the use of multiple displays for a common application. Additionally, we will also mention projects that investigate the position of displays towards users' performance on accomplishing specific tasks. Even though those projects usually draw conclusions on specific tasks that are not related to uCom, they often raise questions that are very relevant to uCom's design. Some of these related projects are detailed next.

3.4.1 Multi-display rooms

The use of multiple displays is common in interactive rooms or smart meeting rooms, which are interactive workspaces where technology is seamlessly integrated into the physical environment. These systems focus in groupware and collaboration. The environments are usually populated with (1) various kinds of displays or projected areas, varying from size - a few inches to full walls; position in the room - on whole or parts of walls, tabletops, on personal mobile devices, chairs, etc; and (2) user interface - mostly composed of touchscreens, pen-based, keyboard and mouse that often allows its cursor to move between displays. Some very well known projects are Fraunhofer IPSI's Roomware [45], and Stanford University's iRoom [28]. Other projects, like Nacenta et al.'s [35], explore how to create a common shared display space from a variety of devices: tabletops, projected surfaces, monitors, and tablets. They address interactions with different display sizes that are not aligned in a single plane. The outcomes of this work are proposed solutions for user interactions on how to navigate between different displays that create a perspective effect from the user's viewpoint.

Some multi-display rooms are set up specifically to experiment the effects of several displays on how users perform specific tasks. Wigdor et al. [55] evaluate the impact of display position while a subject uses an input device to perform a specific visual task on screen. The work explores the impact on user's performance, and any particular user preferences, regarding the display position and user's orientation towards it. It focuses on what mapping of pointing-device input to on-screen movement should be employed for any given screen position. This work is relevant in that it concludes about where to best position the display respect to a user, while the user is directly interacting with it. The experiments provide evidence that participants try to optimize comfort rather than performance. The study also discusses layout guidelines for collaborative computing environment with one or more shared displays, specifically on input-output spatial mappings. Another related work by Plaue et al. [40] investigate how the presence and location of multiple shared displays change the performance and dynamics of teams while collaborating on a data-intensive task. They conducted a study evaluating different positions of a set of shared displays, and concluded that specific locations significantly impacted the ability of teams to make logical connections amongst data.

3.4.2 Spatially Immersive Displays

Spatially Immersive Displays exist in many configurations. One of the most popular is the CAVE^{TM}(Cave Automatic Virtual Environment) [2], a multi-person immersive virtual reality environment invented in 1991. Graphics are projected in this room-sized cube, with high-resolution 3D video and audio, and viewed with active stereo glasses equipped with a location sensor. As the user moves within the display boundaries, the correct perspective is displayed in real-time to create an immersive experience. It is noteworthy that systems like the CAVE^{TM}and some of the interactive meeting rooms require a dedicated setup, regarding the room and its equipment.

Another relevant example is the already mentioned "The Office of the Future" project [41]. It envisions the creation of a life-like shared-room experience by using real surfaces of an office as spatially immersive displays for shared telepresence and telecollaboration.

3.4.3 How uCom deals with multiple displays

Unlike most of the aforementioned project, uCom doesn't require specialized hardware. We aim to incorporate fairly any displays to regularly used real-world spaces, specifically workspaces. It explains uCom's focus on off-the-shelf equipment, and why it does not require specific positioning of equipment throughout the environments. Our vision is to minimally interfere on how a space and its furniture are laid out. For this reason, our ability to provide immersion to a remote space has its limitations, which directly relate to the kinds of devices available and how motivated the user is to feel immersed in the other space. Therefore we are aware that uCom cannot provide the same sense of immersion a user experiences with systems like the CAVE^{TM}.

We also acknowledge the discontinuity inherent in our proposed multi-display interaction. Often times, multiple displays or multiple projection surfaces are positioned in a grid so as to represent an enlarged displayable surface, as aligned images tend to be easily perceived by users as one single large display. It's not the case with uCom, which has to deal with the heterogeneity intrinsic to variable display sizes, irregular positioning, and different kinds of equipment. Our challenge is to create interactions that are suitable to all these various display types.

Concerning user's position respect to multiple displays, uCom tries to address it from two perspectives: (1) how users directly select what is shown in each display, and (2) how to position the displays, which are already showing the chosen views of the remote space. First, our interaction design targets simplicity, providing an easy to use interface to control what image should be shown in each display. For this reason, we proposed a remote control based user interface, that lets users point directly to each display and switch between available images . More details are described in Chapter 4. Second, concerning positioning the displays, in uCom we try to empower the user to position and assign specific views to each display by providing relevant information about the remote space. We argue that, by providing users with a 3D image-based model of the remote scene, we are empowering them to understand the remote space's spatial arrangement, and to locate the remote views that are relevant to a user's particular interests. In addition, the user is also free to position the available displays on locations that can attract appropriate attention without being disruptive. Assuming that uCom is always on, we foresee that users will have the displays positioned on the periphery of their attention, as images of the remote space might not be relevant to their main task at all times.

In the current stage of development uCom is still not directly addressing agreement issues between multiple-participants physically present in each uCom room, i.e. what and where the views of the remote spaces should be displayed. Yet we plan to evaluate in a near future how displays should be positioned in order to optimize their use by multiple participants. As observed by Wigdor et at. [55], in environments where input devices might be shared by multiple, disparately oriented participants, such as table-centered environments, care should be taken to allow participants to make input to any ancillary displays at a desirable orientation. In fact, multiple collocated displays are often used collaboratively, notably in war rooms and operating rooms. In those environments, users and their input devices are often not located directly in front of, or oriented towards, the display of interest. Given that in real environments it might not be possible to position displays and orient control spaces to satisfy each user's preference, we should evaluate in future user studies the penalty on performance if either or both user's preferences are not met. How willing are users to engage with displays if they are not in individually owned? Will they have lower frequency of use if they are not individually owned?

Nevertheless, we are aware that other factors are relevant for ambient or peripheral displays, such as: aesthetics, utility, form factor, positioning restrictions, easy to use, integration with user's environment, comprehension of its content, among others.

3.5 Awareness Applications

The definition of workgroup awareness found in the literature is fairly coherent, even though multiple terms commonly refer to similar ideas. In a very relevant work, Dourish & Bly [18] define awareness as knowing who is around, what activities are occurring, who is talking with whom; providing a view of one another in the daily work environments. They emphasize that awareness may lead to informal interactions, spontaneous connections, and the development of shared cultures. Another similar definition proposes the term general awareness as the pervasive experience of knowing who is around, what others are doing, whether they are relatively busy or can be engaged [22]. It is noteworthy that CSCW researchers have long acknowledged the importance of awareness in facilitating collaborative work [17].

In our work we we are interested in assessing how uCom can convey ongoing awareness, in the sense of passive mutual monitoring of each other's activities between users situated at remote locations. A lot of the research in awareness between remote workgroups has been done by systems that have audio and video communication features. Most of these projects are commonly referred as *media spaces.* Even though uCom does not enable audio communication, it explores research questions related to awareness in media spaces. Therefore, we will briefly describe media spaces and related projects as follows.

3.5.1 Media Spaces

According to Baecker [10], a media space is "a computer-controlled teleconferencing or videoconferencing system in which audio and video communications are used to overcome the barriers of physical separation". Media spaces focus on providing awareness to foster collaboration, or supporting two or more people while intensely focused on a common task. The idea is to support physically separated colleagues to work naturally together using media spaces. It is worth mentioning some classic media space projects. First, we must mention Xerox PARC's Media

Space [46], which originated the Media Space term. It started in 1985, as an environment for cooperative design, inspired in open architectural studios. Other relevant projects are EuroPARCs RAVE [22], similar to Xerox PARC's Media Spaces, but connecting multiple users in the same building, providing general awareness, and often leading to serendipitous communication. The system was used on a daily basis, creating a distributed community. It's one of the first relevant experiments of this kind, exploring norms of privacy, connection management focused on intrusion prevention, work culture and technical constraints.

During the 1990's, the growth of the Internet increased the use of media spaces to enable collaboration among remotely distributed colleagues. Other classical media space projects are mentioned as follows: Portholes [18], Polyscope [13], CAVECAT [33], Bellcore's VideoWindow and Cruiser [15] [21], in addition to the MIT Media Lab's project that initially inspired this work: iCom [6] [8]. Those projects explored video and audio connection in offices and common workspaces or even for always-on office-office connections. A common focus was supporting casual or informal communication, and to mimic physical proximity. Most of them explored related issues such as privacy, unobtrusive awareness, informal interactions, collaboration, control and symmetry.

3.5.2 Spatial models and remote collaboration

A particularly interesting aspect of media spaces is the creation of communication modes appropriate to specific situations. For instance, EuroPARC's RAVE [22] segmented connections in three distinct modes: "glance" - one-way, short-term, video-only; "video-phone" - two-way, long-term, video and audio; and "watch" - one-way, long-term, video and audio. Each of those connection modes could be configured with different accessibility levels. The Montage project [50] also focused on providing audio-video glances among distributed team members. The glance was analogous to peeking into someone's office to check his availability. The goal was to increase the accessibility to individual, without disrupting them. The metaphor was “to open someone's office door” to assess the person's availability using video and audio.

On the other hand, there has been lots of criticism on physically imitating face-to-face communication. Hollan and Stornetta [26] argue that most media spaces are too focused on imitating face-to-face communication. They question if any technology will be ever powerful enough to make those at distance at no real disadvantage to those collocated. Their proposal is that, instead of trying to mimic physical proximity, telecommunication research should develop tools that go *beyond being there*, i.e., tools that people would prefer to use even when the possibility of physically interacting is available. For that, they propose framing human communication problem in terms of three aspects: (1) need - human requirements that encourage and facilitate interaction, (2) media - the medium that enables communication, and (3) mechanism - ways to meet informal communication needs through the medium. The framework's goal is to identify needs which are not ideally met while in physical proximity, and therefore propose mechanisms to enable the medium to meet those needs. Dourish [16] also criticizes the approach of spatial models and metaphors in collaborative systems. He claims that designing collaborative systems that mimic the spatial organization of the real world is too simplistic. Indeed, several Telepresence systems have attempted to simply replicate features of spaces expecting to enable behaviors that are appropriate in their real-world counterparts. Dourish argues that the notion of "place" is what actually frames the interactive behavior, socially and culturally, not the technology *per se*. CSCW tools create new social places when users attribute social meanings to new technological features.

Another common argument is that awareness applications should be matched with appropriate tasks. Often times those systems are evaluated according to user's performance on a specific task. For instance, Gaver et al.'s [23] studies suggest significant constraints due to the limited field of view of a single fixed camera in most remote collaboration applications. Their evaluations indicated that participants preferred task-centered rather than face-to-face or head-and-shoulder views to collaborate in specific task.

An early 1990's paradigm of the typical media space node - a video monitor, a camera, a microphone and nearby speakers - is still very popular in current end-user videoconferencing systems like Skype^{TM}[2], which also allows users to maintain "always-on" video and audio connections. Yet software like Skype are usually focused on solely portraying one remote individual at a time, limiting the visualization of other aspects of the remote environment, such as the whole space's arrangement, other people, and any other events that might be taking place beyond the camera's field of view.

3.5.3 How uCom relates to awareness systems

uCom cannot be considered a media space, as it does not provide audio communication. Yet, similarly to media spaces, uCom also focuses on conveying ongoing awareness. In uCom's case, awareness relates to passive mutual monitoring of each other's activities between users situated at remote locations. The awareness mode that best describes uCom's features is the "glance", a one-way, short-term, video-only into a remote space.

Indeed, media spaces are often criticized for an excessive focus on imitating face-to-face interaction or sharing views of particular artifacts relevant to specific tasks [26] [23]. On the other hand, uCom focuses on portraying the remote space itself, not its specific users. Multiple live video views of the remote space can be browsed in a 3D image-based model that provides a sense of its scene geometry. Differently from most media spaces, uCom doesn't create a "space" for direct interaction. It exclusively provides visual awareness alongside the everyday physical environment.

While most media spaces are evaluated with respect to user's performance in specific collaborative tasks, uCom's evaluation targets the visual perception of the remote space *per se.* Even though uCom does not provide collaboration features, it enables group engagement by creating a shared context through its visual live video views embedded in the physical environment.

[2] Skype: http://www.skype.com

4

uCom Prototype

In this chapter we describe the implementation of uCom's prototype. We present an overview of the system architecture, including its hardware and software requisites. Next, we present the system configuration and steady state operation cycles, followed by detailed description of the software subsystems.

4.1 System Architecture Overview

uCom's system architecture combines a set of hardware and software that enable multiples live video views and scene information to be mutually transmitted between two uCom rooms. uCom's hardware setup is composed of multiple cameras and displays. One uCom room can show on its displays two possible display formats from the other room: any individual video views, or multiple video views assembled in a 3D representation of the remote scene, which resembles a 3D collage. uCom's software implementation has two main subsystems: (1) the 3D scene image-based model computation and rendering subsystem, and (2) the selection of remote views to local displays' subsystem.

The purpose of the first subsystem is to create a 3D image-based model of the scene, which can be navigated by its users. It uses as inputs video feeds acquired from multiple viewpoints of a remote scene to render a visual representation based on the estimates of the scene and cameras' geometry provided by Bundler [1], a software designed by Snavely et al. [42] [43] [44]. More details on how Bundler works are provided further on this chapter. Additionally, Section 4.4 presents some examples of 3D image-based models of representative scenes.

The second subsystem allows users to determine which live video views from the remote space should be portrayed at each display that is locally available. For that, users utilize a simple remote control interface. More details on this subsystem are provided further on this chapter.

It should be noted that the current uCom prototype doesn't focus on any specifics of the following networking aspects related to transmitting multiple live video streams, such as: synchronization, delays, jitter, error resilience, or quality of service. We are also not directly addressing uCom's computational complexity regarding 3D graphics support and the number of video streams the system can handle. Our current focus is to show that the system is feasible in specific conditions. More generalized operational conditions might be addressed in future work.

4.1.1 Hardware

The ideal system's physical setup is composed of two uCom rooms, each one with multiple cameras and multiple displays. One uCom room can show on its displays two possible display formats from the other room: any individual live video views, or multiple live video views assembled in a 3D representation of the remote scene, which assembles live video feeds acquired from multiple viewpoints of the scene in a graphical visualization resembling a 3D collage. Figure 4-1 shows that multiple images from one room can be displayed in multiple displays in another room, and vice-versa. The correspondence is **m to n**, i.e., **m** cameras to **n** displays. All video and image data between two uCom rooms is transmitted over a network connection, and the user interface with the system is through keyboards and remote controls. It should be noted that cameras and displays are connected to computers and input devices, even though they are not explicitly represented in some of our figures throughout this chapter. All software components run on Apple$^{\mathrm{TM}}$Intel Mac computers, mostly Mac Minis [1] running Apple$^{\mathrm{TM}}$Mac OS X 10.5[2] operating system.

[1] Apple Mac Mini: http://www.apple.com/macmini/

[2] Apple Mac OS X 10.5: http://www.apple.com/macosx/

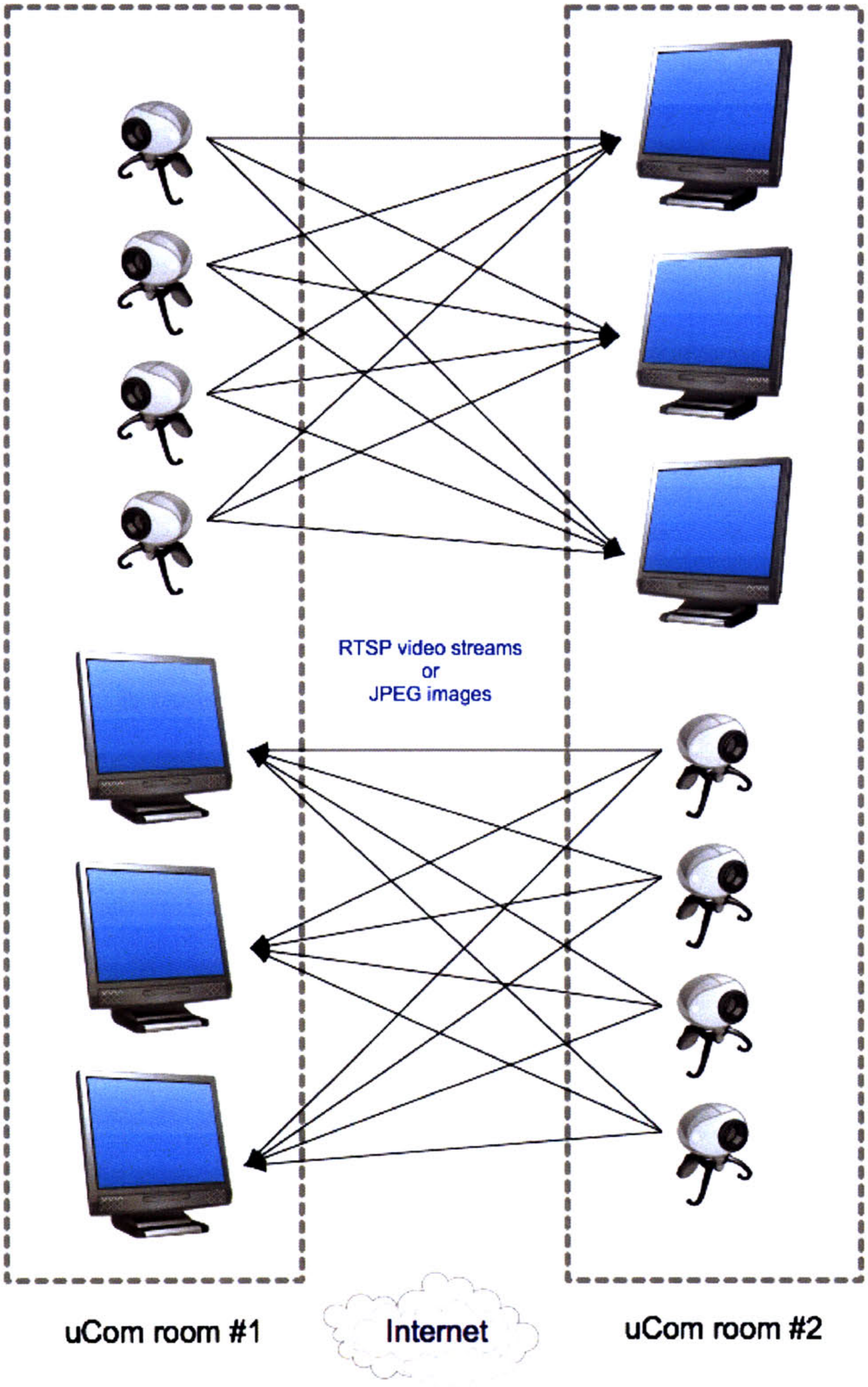


Fig. 4-1 Correspondence from m cameras to n displays

As per the user interface, the keyboard enables users to navigate the 3D image-based model of the scene, while the remote control is used to select which images should be shown in each of the displays.

4.1.2 Software

uCom's software is implemented using C and C++ programming languages, and additional libraries that are mentioned further on. The system implementation currently comprises two independent subsystems: (1) one that computes the 3D image-based model of a scene, and (2) one that allows users to enable remote live views on local displays. They should be integrated in a future implementation, but at present they operate independently. The two subsystems are computed in distinct cycles. First, the 3D model of the scene is computed only during the configuration cycle, but users can navigate the 3D model after configuration has been completed. The other subsystem allows users to switch between remote views of a uCom room on local displays at all times. Figures 4-2 and 4-3 show diagrams of the uCom subsystems. But it is important to clarify that the system portrayed in the aforementioned figures is symmetric, as both uCom rooms have the ability to acquire and display live videos.

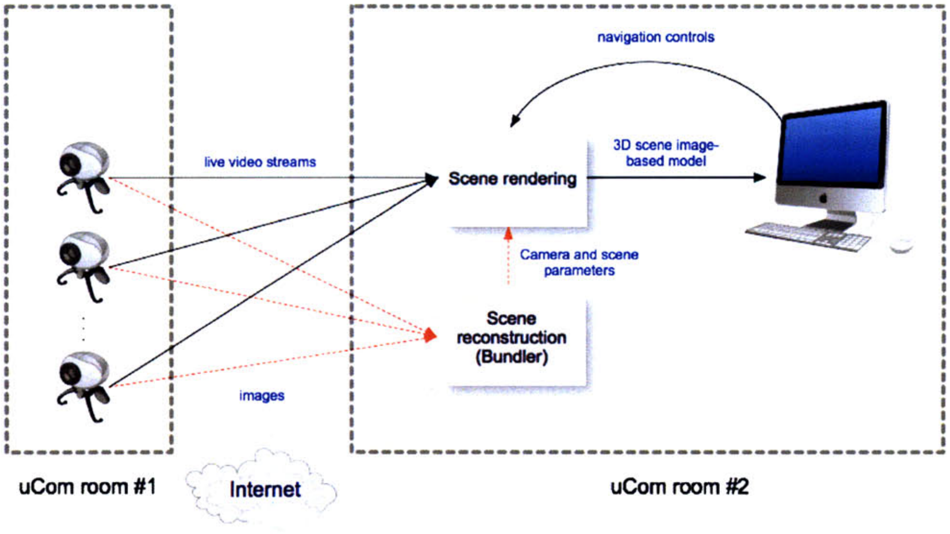


Fig. 4-2 uCom subsystem: 3D scene image-based model computation and rendering.
Red dashed lines refer to the scene configuration cycle, while black continuous lines refer to the steady state cycle.

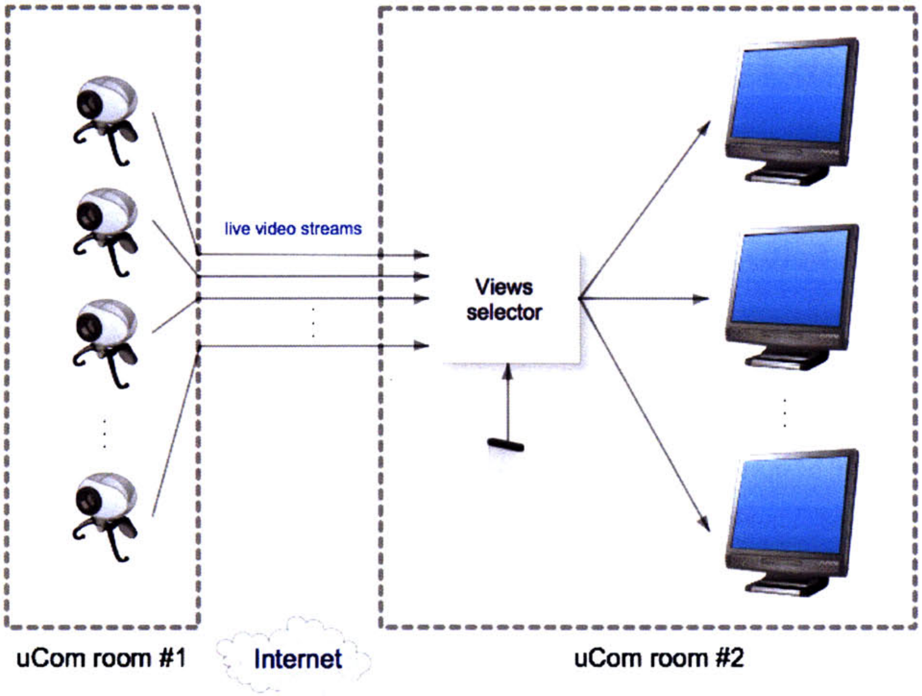


Fig. 4-3 uCom subsystem: selection of remote views to local displays

The system operates in two distinct cycles: configuration and steady state. The configuration basically refers to positioning cameras, taking snapshots of the scene and computing a 3D image-based model of the scene. After this computation, all cameras are expected to remain stationary, and the system can operate continuously. Every time any of the cameras is moved, the system needs to be re-initialized.

4.1.3 Configuration cycle

The configuration process takes snapshot photos of the scene in JPEG format, and inputs them to Bundler [1], which takes this unordered set of images, matches them and outputs an estimate of a sparse 3D model of the scene geometry. The output is specifically composed of: (1) feature points' 3D position in the scene, and their 2D position in each image they were detected, and (2) cameras' position and orientation. The scene configuration cycle is represented by the red dashed lines in Figure 4-2. Section 4.2.2 describes how Bundler works.

Concerning uCom's image capturing process, it should be noted that the images need to overlap considerably so that the scene can be sufficiently reconstructed. This means that scene features should be visible in at least two images. According to Snavely et al [44], a ῗule of thumb" for Bundler's computation is that a pair of neighboring images must overlap by more than 50%, which means that at least half of the area portrayed in one image must appear in another image. Additionally, we recommend that the images of the scene should be acquired while nobody is present in the room, so as to prevent people from occluding static features of the scene.

4.1.4 Steady state cycle

After the system is initialized, we consider it operating in a steady state cycle, in which the cameras remain stationary so that the 3D image-based model can maintain correspondence to the position from which each video is acquired.

During the steady state cycle, uCom's user interface enables two kinds of interactions: (1) using the keyboard to navigate the 3D image-based model of a remote uCom room, and (2) using a simple remote control to select which of the available views of a remote space should be shown in each of the displays locally available. More details on the user interface are presented in each subsystem's description.

We use OpenGL transparency blending features to enable a general sense of how the multiple video views match together in a 3D collage. It should be noted that only one video view is selected by the user at any one given time. We want to emphasize the selected video, while giving a sense of what lies outside its field of view. So, we render the selected video as opaque and in the foreground, while the non-selected views are rendered as translucent or partially transparent. Section 4.4 presents some examples of 3D image-based models of representative scenes.

We implement additional features to allows users to navigate the 3D scene using keyboard controls implemented with libSDL. The 3D transitioning between live video views is rendered so as to preserve the spatial relationship between the views. The transition navigation controls are listed as follows:

4.2 3D Scene Image-based Model Computation and Rendering Subsystem

The purpose of this uCom subsystem is to acquire videos from multiple viewpoints of a remote scene and render a 3D image-based model of the scene, which can also be navigated by its users. The goal is not to create a photorealistic view of the scene, but to display live video views in spatial context, providing a sense of the geometry of the real scene. This subsystem is implemented with the integration of multiple software components. (1) a video stream client, (2) a scene reconstruction algorithm, and (3) a rendering mechanism with navigation controls. Some of the components were developed from scratch, like the scene rendering mechanism and scene navigation controls; while the video stream client and scene reconstruction components were integrated from existent algorithms and libraries. We describe each of these components in the following subsections.

As previously mentioned, the scene reconstruction algorithm is only computed during uCom's configuration process. Once the scene is estimated, the system runs in steady state considering as static both the camera pose and the scene geometry. The scene reconstruction computation can be refreshed in case the cameras are moved or the scene changes considerably, but we haven't implemented this updating feature in the current prototype.

4.2.1 Video streaming client-server

All video data captured by cameras located in one uCom room must be delivered through the network to another uCom room, where they are displayed. We transmit these video streams using RTSP - Real Time Streaming Protocol. We use video streaming servers to capture live video from the cameras, and video streaming clients to receive the videos then send them to uCom's rendering component to be shown on a display.

On the client side, uCom makes use of a reliable multi-platform video streaming client, VideoLAN VLC media player [3], available via GNU General Public License [4]. This media player is integrated to our system architecture using libVLC [5] library. Its output is relayed to an SDL [6] pipeline, which converts each video frame to an Open Graphics Library (OpenGL) texture.

On the server side, uCom supports any video stream servers that can multicast using RTSP. In our current sets of experiments we use VideoLan VLC media player or Apple QuickTime Broadcaster^{TM}[7].

It should also be noted that VideoLAN VLC media player can seamlessly load MPEG-4 files rather than the RTSP video streams, which can be useful for testing purposes.

4.2.2 Scene computation using Bundler

uCom makes use of an estimated scene and camera geometry computation provided by Bundler [1], designed by Snavely et al. [42] [43] [44]. But unlike Snavely et al., we are not interested in using the plethora of photos of outdoor scenes available on the Internet. uCom targets videos of bounded scenes, indoor or outdoor. Yet, in current user studies we focus on indoor office spaces.

It is important to clarify that uCom uses Bundler to "register" the video views in space, which means to estimate the correspondences between images, and how they relate to one another in a common 3D coordinate system. Bundler is used

[3] VideoLAN: http://www.videolan.org

[4] GNU GPL: http://www.gnu.org/licenses/

[5] VideoLAN libVLC library: http://wiki.videolan.org/Developers_Corner

[6] Simple DirectMedia Layer libSDL library: http://www.libsdl.org/

[7] Apple QuickTime Broadcaster: http://www.apple.com/quicktime/broadcaster/

during uCom's configuration process to compute the 3D image-based model of the scene that will be used on uCom's steady state cycle. More details on the configuration process were mentioned in Section 4.1.3.

uCom takes advantage of Bundler for its scene reconstruction for two main reasons. First, Bundler can robustly handle heterogeneous inputs: unordered images, taken from many different cameras, with variations of lighting, etc. Second, all of Bundler's computations rely exclusively on computer vision, not requiring any special hardware.

Snavely et al. have implemented both scene reconstruction and scene visualization features, but only the scene reconstruction algorithm is distributed under the GNU General Public License. Therefore, uCom only uses the scene reconstruction pipeline implemented by Snavely et al. in Bundler [1], but not the scene visualization algorithms of Photo Tourism [42] [43] [44] and Microsoft Photosynth [7]. We have implemented uCom's own scene visualization and rendering features following methods suggested in Snavely's PhD dissertation [44]. Additionally, we have implemented a system that renders video frames rather than stationary images.

Bundler system architecture

Bundler's reconstruction estimates the camera geometry - its orientation, position and internal parameters, as well as the scene geometry - a sparse set of 3D feature points, that represent salient points belonging to objects in the scene. Bundler takes as inputs a set of of images of a scene, which must have a considerable overlap (see Section 4.1.3), and outputs a set of parameters of the reconstructed scene model and cameras. Details on Bundler's inputs and outputs are presented below.

- Cameras: estimate intrinsic and extrinsic camera parameters. For each camera C_j:
  - focal length scalar f_j
  - two radial distortions scalars K1_j, K2_j

  - a 3x3 matrix R_j representing the camera orientation (rotation respect to the system base coordinates)
  - a 3-vector t_j describing the camera location (translation respect to the system base coordinates)

- Points: estimates of points in the real world scene. For each point P_i:
  - 3-vector describing the point location (translation respect to the system base coordinates)
  - a list of cameras from which the point is visible:
    - camera
    - index of the SIFT keypoint where the point was detected in that camera
    - (x,y) the detected position of the keypoint in that camera image

The main steps in Bundler's computation are detailed as follows.

Correspondence estimation between images

The first stage of Bundler's pipeline finds correspondences between the images. It uses well known computer vision techniques to identify which pixels in different images correspond to the projection of the same 3D point in the real world. Those techniques mostly rely on the assumption that the same 3D point in the world has similar appearance in different images, especially if the images were taken from not so distant positions. The main steps computed on Bundler's correspondence estimation follow.

1. **Feature detection**: Feature points are detected using Scale-Invariant Feature Transform (SIFT) [31], which finds image features by applying a differences of Gaussian (DoG) filter to the input image. Each (x,y) position of a maximum or minimum output by the filter is considered a feature. SIFT is considered well suited to matching unordered images at different resolutions and points of view, as it is very robust to image transformations, such as

variances of scale (how zoomed in an object is in different images), brightness and contrast. SIFT outputs both the feature location and a descriptor, which is a vector describing the image gradient around the location of the feature.

2. **Feature matching**: For each pair of images, its respective SIFT keypoints are matched using an Approximate Nearest Neighbor algorithm implemented by Arya et al. [47]. A distance metric between two matching candidates is computed using Lowe's ratio test [31], refined with a threshold. After removing spurious matches, each feature from image I_i will have at most one matching feature from Image I_j. If Images I_i and I_j have too few matches, all SIFT points are eliminated, and the images are not considered to match. It is noteworthy that this matching procedure is imperfect. So further spurious matches pass through a geometric consistency test, which eliminates sets of matching features that are not physically realizable according to each two cameras corresponding epipolar geometry. The epipolar constraint between two cameras is defined by a *fundamental* matrix, which is computed with a RANSAC model-fitting algorithm [20]. After computing the feature matching between all $\binom{n}{2}$ image pairs, those matches are grouped into *point tracks*, which are list of matching keypoints across multiple distinct images that represent the 3D location of the same point in the real-world.

Scene reconstruction or recovery

The second stage of Bundler's pipeline aims to simultaneously reconstruct the camera parameters and the 3D location of the *point tracks.* It uses Optimization techniques and Structure from Motion (SfM) to compute the configuration of cameras and 3D points that best match the feature correspondences when related through perspective projection equations. Those techniques take advantage of the fact that if we have enough point matches between images, the geometry of the system becomes constrained enough to determine the camera poses.

Bundler's SfM approach is similar to that of Brown and Lowe [14], with some modifications to improve robustness over varied data sets and maximize its likelihood of success, which the main ones are: (1) using a differentiated heuristic to select the initial pair of images for SfM, (2) checking if reconstructed points are well conditioned before considering them part of the scene estimation, (3) initializing the computation with focal length estimates available from the JPEG input image's EXIF tags. Those modifications are detailed as follows.

In order to prevent the optimization algorithms to get stuck in bad local minima, good initial estimates for the parameters are needed. Therefore, Bundler focuses on incrementally estimating parameters for pairs of cameras that have a higher likelihood of success. It does so by initializing the system with a pair of images that have at the same time a larger number of keypoint matches and a larger baseline. It also makes sure the cameras cannot be well modeled by one single homography, which prevents the degenerate case of coincident cameras.

Then, in each optimization iteration, a new camera is added to the computation. The newly selected camera must observe the largest number of keypoints from the tracks whose 3D position have already been estimated. The new camera parameters are also initialized: the extrinsic parameter are obtained with a direct linear transform (DLT) [25] inside a RANSAC computation, and the intrinsic parameters are started with estimates from the image file's EXIF tags.

Afterwards, tracks are added to the optimization if two conditions apply: (1) its respective feature points are observed by at least one other already recovered camera, and (2) if triangulating its SIFT feature position in both camera sensors' current estimated positions returns a well conditioned estimate of its scene point's 3D position.

The optimization procedure is repeated, reconstructing the scene incrementally. It adds one new camera at each iteration, until no remaining cameras can observe any of the reconstructed 3D points. The objective function is minimized at each iteration using a modified version of the Sparse Bundle Adjustment package by Lourakis et al. [30] as the underlying optimization engine.

The algorithm described above was later changed to improve its speed and robustness. First, after each optimization run, outlier tracks are removed. The removal condition is having at least one feature keypoint with higher reprojection error. Second, multiple cameras are added at a time in the optimization, instead of just one. The heuristic adds at each iteration all cameras with at least 75% of the maximum number of matches to the already recovered 3D point.

Further details about the aforementioned algorithms can be obtained from Noah Snavely's dissertation [44], related papers [42] [43], and Bundler's website [1].

4.2.3 Rendering

This software component uses the scene reconstruction parameters computed by Bundler [1] to render live video views from a remote uCom room. It renders them in an interactive 3D image-based model, which resembles a 3D collage that is faithful to the scene geometry. Section 4.4 presents some examples of 3D image-based models of representative scenes.

The Rendering component performs three main operations: (1) it computes an estimate of where the live video views should be rendered using scene reconstruction parameters generated by Bundler, (2) it renders the 3D image-based model of the scene from specific viewpoints, and (3) it creates input controls that enable users to navigate the 3D model of the scene.

It should be noted that uCom's current prototype does not make use of any scene visualizations software implemented by Noah Snavely's on his PhotoTourism Project [42] [43], or Microsoft Photosynth [7]. Even though Snavely et al. have implemented both scene reconstruction and scene visualization features, only Bundler [1], the scene reconstruction algorithm, is distributed under the GNU General Public License. Therefore, uCom only uses the scene reconstruction pipeline implemented by Snavely et al. in Bundler [1], but not the scene visualization algorithms of Photo Tourism [42] [43] [44] and Microsoft Photosynth [7]. Yet, we have implemented uCom's own scene visualization and rendering features following methods suggested in Snavely's PhD dissertation [44]. Additionally, we have implemented both a system that renders video frames rather than stationary images, and our specific navigation controls described further on.

Image position estimator

This component computes the 3D positions to which the images of the remote scene should be rendered in order to create a 3D collage effect. It uses as inputs the *a priori* knowledge of the static scene's geometry estimated by Bundler, which is composed of the cameras' parameters, 3D feature points of the scene, and one additional parameters: the scene estimated up vector. Our heuristic estimates the 3D position of a quadrilateral to which each image should be rendered. The two main steps of this process are: (1) computing the 3D plane to which an image should be projected, and (2) estimating, on this 3D plane, the position of the quadrilateral area to which the image should be rendered.

First, in order to compute the 3D plane to which an image should be projected, we use two sets of data: (1) the positions of the 3D feature points that Bundler identifies as visible from the camera that has acquired this specific image, and (2) the estimated up vector of the scene. Our algorithm projects the images onto a plane that is approximately vertical respect to the ground plane. This plane is computed by projecting the 3D feature points onto the scene's estimated ground plane, fitting those projected points into a line, and raising a plane that contains this line and that is also perpendicular to the scene's estimated ground plane. The scene "up" or gravity vector is computed using the method proposed by Szelisky [48], which is already implemented in Bundler's original source code. This heuristic assumes that the real world scene has a dominant plane that is perpendicular to the estimated scene ground plane. The idea is based on the assumption that many planar surfaces that occur in the world are walls or the floor. This approximation is also used in some of Snavely et al.'s experiments [42] [43] [44]. It works particularly well with photos of outdoor touristic sites used as examples in their work, as they usually portray building facades, which have most of its feature points approximately on one plane that tends to be perpendicular to the scene ground pane. As uCom focuses on indoor scenes, it is likely that feature points are detected on both the walls and on foreground objects like furniture. In this case, it is probable that projecting the image onto one single plane does not work that well with indoor scenes as it does with building facades. More details on these differences are discussed in Section 4.4.

Second, in order to compute the image corners on the estimated 3D plane, we project onto this plane each corner of the respective camera's image sensor's 3D position. We use the estimated camera position and pose to project light rays that connect each of the sensor's corners to the camera focal point. We extend this light ray line till it crosses the previously estimated 3D plane to which the image should be projected. These intersections between light rays and projection plane determine the corners of the quadrilateral to which each image should be rendered. Figure 4-4 shows this projection procedure.

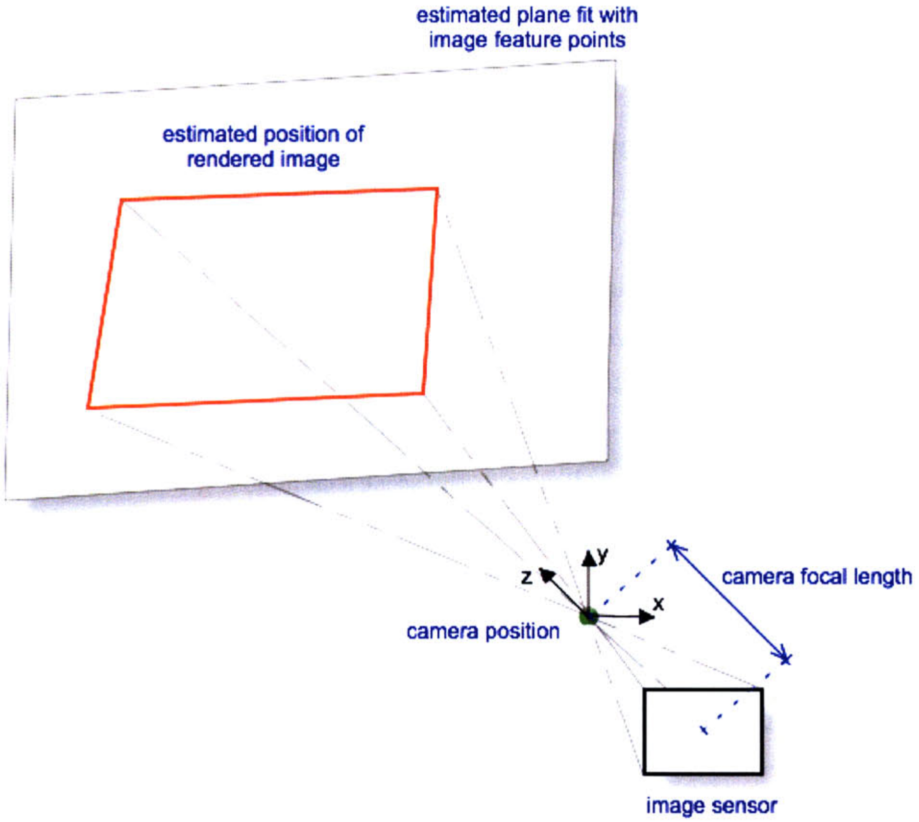


Fig. 4-4 Rendered image's position

In summary, our algorithm computes a quadrilateral area to which each image should be rendered. A more detailed description of the computational steps is presented below.

For each camera C_i, and its respective image Img_i and 3D feature points FP_i:

- project all 3D feature points FP_i onto the scene estimated ground plane,
- fit a 2D line l_i to the points projected on the ground plane,
- raise a vertical plane P_i containing the 2D line l_i,
- compute the 3D points that delimit the quadrangular area in plane P_i to where the image Img_i should be projected.
  - As show in in Figure 4-4, for each corners of camera C_i's image sensor:
  - compute the line equation that connects the image sensor's corner to the camera position (or center of projection),
  - compute the position where this line intersects plane P_i.

Scene Renderer

uCom requires a visualization front-end that can render live video streams rather than stationary images. For that we implement our own scene rendering mechanism. We render the video frames using the image position estimation methods detailed in the previous sections. Additionally, we implement features that allow users to navigate the 3D scene by moving from one camera viewpoint to another, zoom and pan.

As previously mentioned, our software integrates three libraries: (1) libVLC, which enables access to VideoLAN VLC's video streaming client, (2) libSDL, which handles the user interface (window building and keyboard inputs), and also accesses the video buffer, converting each video stream's frame to an OpenGL texture, and (3) OpenGL, which renders all videos on a 3D graphical user interface.

We use OpenGL transparency blending features to enable a general sense of how the multiple video views match together in a 3D collage. It should be noted that only one video view is selected by the user at any one given time. We want to emphasize the selected video, while giving a sense of what lies outside its field of view. So, we render the selected video as opaque and in the foreground, while the non-selected views are rendered as translucent or partially transparent. Section 4.4 presents some examples of 3D image-based models of representative scenes.

We implement additional features to allows users to navigate the 3D scene using keyboard controls implemented with libSDL. The 3D transitioning between live video views is rendered so as to preserve the spatial relationship between the views. The transition navigation controls are listed as follows:

- Switch between camera viewpoints: the scene can be visualized from the viewpoint of each estimated camera. The transitions are rendered smoothly along a straight line between the estimated camera positions, so users can visualize the path between two neighboring views. We cross-fade from the previous to currently selected image.
- Pan and zoom: users can move in any direction along the selected camera's axis (x, y, z), in order to reposition their viewpoint closer or farther, or even to the left, right, up and down to better fit users' needs.

4.3 Selection of Remote Views to Local Displays' Subsystem

This independent subsystem allows users to determine which live video views from the remote space should be portrayed at each display that is locally available. This way, we try to extend the vision of "spatially immersive displays" onto regular screens. Furthermore, our design focuses on simplicity. So we create a very simple user interface with which users can easily assign a remote view to any of the local displays, without being restricted to use a specific computer's mouse and keyboard. Figure ?? shows our proposed solution, containing a simple remote controlled system. This design allows a user to move more freely while deciding which remote views would be appropriate at each display.

We implement this software subsystem using Apple Quartz Composer$^{\text{TM}}$, which enables very quick prototyping with graphical features. Our system allows users to assign each display with one single video view of the remote space using a simple Apple remote control$^{\text{TM}}$(Figure 4-5). The "left" and "right" buttons are used to select the display to which a user can assign a different remote view. By pressing "left" or "right", a user moves a visual beacon between displays, which indicates the currently selected displays. Once the desired display shows the visual beacon, a user can use the "up" and "down" buttons to change the remote view to be shown on its screen. The user repeats this process untill all available displays in the uCom room had been assigned a remote view that fits user's needs. It should be noted that the list of remote views or local displays, sorted with "up-/down" and "left/right" buttons respectively, is currently a circular list determined on uCom's configuration process. Users can choose the ordering of remote views and local while initializing uCom, e.g., displays that are physically side-by-side can be initialized with consecutive numbers.

Fig. 4-5 Apple remote control$^{\text{TM}}$used to assign remote views to local displays

4.4 Sample 3D Image-based Model of Selected Scenes

In this Section we present samples of selected scenes assembled in a 3D image-based model, in order to exemplify how this graphical visualization looks like. For that, we execute Bundler and our scene rendering and visualization algorithm using as input about 10 images of each selected scene. Then, we render only a subset of the views that provided the most visually compelling 3D image-based model.

Next, we present two sets of selected scenes that are representative of use case scenarios we foresee to uCom's current prototype. They are (1) a bulletin board wall, and (2) a wall and some furniture placed very close to it. The latter was acquired in three different situations: (1) using stationary images acquired by camera sensors approximately parallel to the wall, (2) using stationary images acquired by camera sensors in different orientations with respect to the wall, and (3) using video feeds acquired by camera sensors approximately parallel to the wall.

The aforementioned scenes were selected in order to exemplify two main aspects of the 3D image-based model computation. First, images acquired while the camera sensor is parallel to a wall don't generate projection distortions, as the quadrilateral area to which the image is rendered is approximately a rectangle. Second, scenes in which foreground objects - like pieces of furniture - are very close to the background wall tend to generate a satisfactory 3D image-based model. This is due to the fact that the algorithm tries to estimate one single plane onto which the image should be rendered that contains feature points belonging to both foreground objects and the background wall. Therefore, if foreground objects are very close to the wall, the algorithm estimates a plane that is much closer to the real-world scene than if foreground objects are too far from the background wall.

Additionally, it should be noted that our visualization uses transparencies to"fade out" all images besides the one that corresponds to the selected camera. It renders all other images with respect to the selected camera's viewpoint while highlighting the image that was captured by the selected camera.

4.4.1 Bulletin Board

In this example the image sensor was positioned parallel to the wall where the bulletin board is displayed, and all feature points on the scene belong to a background wall. Figure 4-6 shows the original images, while Figure 4-7 shows the images positioned in a 3D collage using the 3D image-based model from the viewpoints of both cameras. It's noticeable that the edges seen in both images are closely matched in the 3D image-based model portrayed in Figure 4-7.

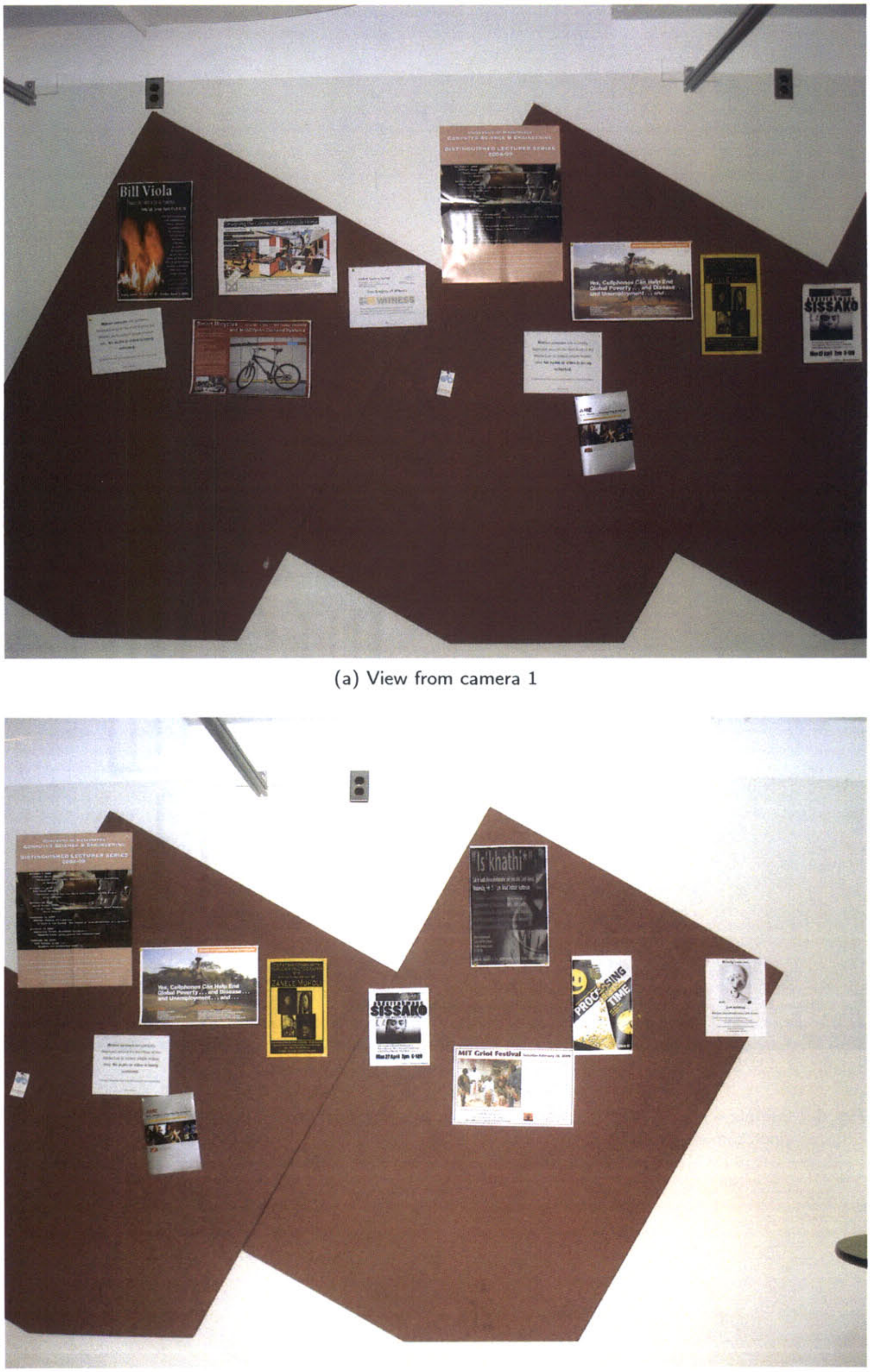
4.4.2 Wall and furniture close to the wall

In these examples, we portray a scene containing a wall and furniture placed very close to this wall. We show the same scene from different viewpoints and image sensor orientations, which are: (1) stationary images acquired by camera sensors approximately parallel to the wall, (2) stationary images acquired by camera sensors in different orientations with respect to the wall. Next we present the case when images are replaced by video feeds acquired by camera sensors approximately parallel to the wall.

First, we show stationary images acquired by camera sensors approximately parallel to the wall. Figure 4-8 show the original images, while Figure 4-9 show these images assembled in the 3D image-based model. The resulting 3D image-based model portrays a satisfactory collage in which most of the edges of windows and furniture are reasonably aligned, providing a good sense on how the separate images fit together onto the scene. Additionally, as the image sensor is parallel to the wall, there are no significant projection distortions, which can be noticed by the fact that most window and door frames seem to be straight in the 3D image-based model.

Second, we show stationary images acquired by some camera sensors approximately parallel to the wall, and only one of them not parallel to the wall (acquired by camera 1). Figures 4-10 and 4-11 show the original images, while Figures 4-12 and 4-13 show these images assembled in the 3D image-based model. It is noticeable in the 3D image based model that the image acquired by an image sensor that is not parallel to the wall (camera 1) is portrayed as a non-rectangular quadrilateral. This view also shows distortions on the edges of its foreground objects, which don't seem straight.

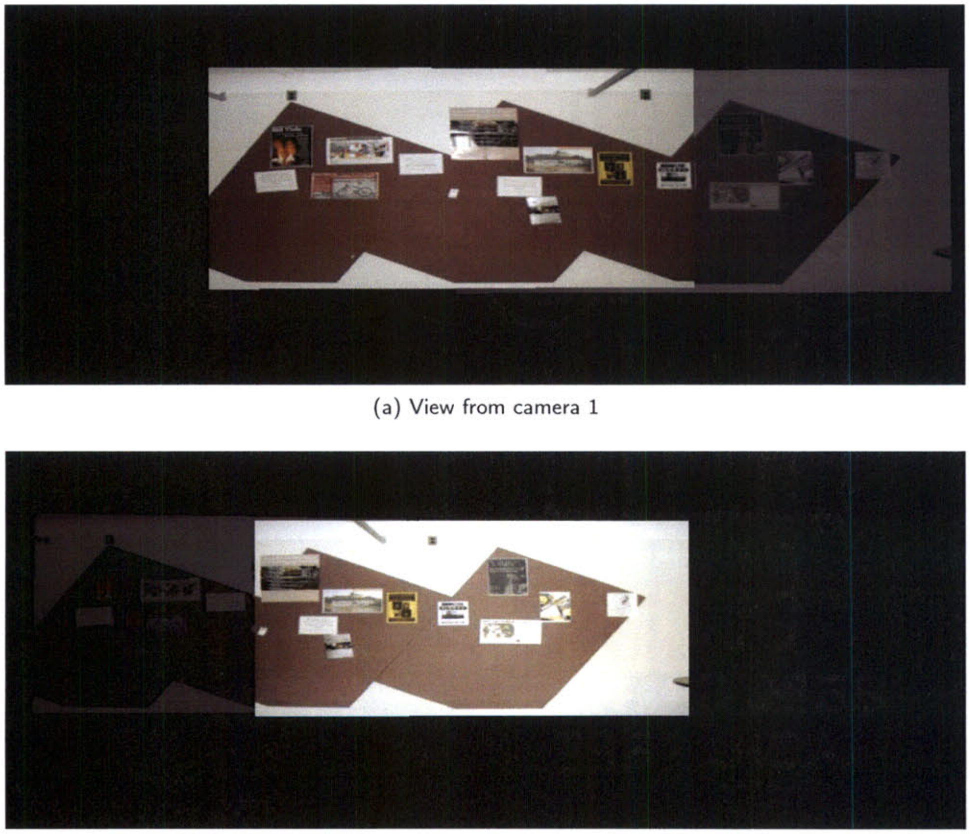
Third, we show three sample frames of a video that records a person walking through the scene from right to left. For that, we use the same cameras' position and orientation as in the previous example, which are shown in Figures 4-10 and 4-11. Figure 4-14 shows the original video frames, while Figure 4-15 shows these frames assembled in the 3D image-based model. It should be noted that in both Figures 4-14 and 4-15, the first two frames are captured from camera 3's viewpoint, and the third video frame is captured by camera 2's viewpoint. The reason why those frames were selected is clearly noticed from Figure 4-15. As the person walks by the scene from right to left, on the first and second selected frames, the person can clearly be seen from camera 3's viewpoint. Yet, on the third selected frame, the person can only be seen from camera 2's viewpoint. The 3D image-based model of the scene shown in Figure 4-15 conveys an interesting behavior on the second frame: as the person is simultaneously seen from two viewpoints, a small part of the person's legs is seen on the background, while most of the person's body is seen on the foreground at a position that do not exactly match with the background image.


Bill Viola
DISTINGUISHED LECTURER SERIES
Yes, Cellphones Can Help End Global Poverty ... and Disease ... and Unemployment ... and ...
SISSAKO
"Is'khathi"
PROCESSING TIME
MIT Griot Festival

(a) View from camera 1

(b) View from camera 2

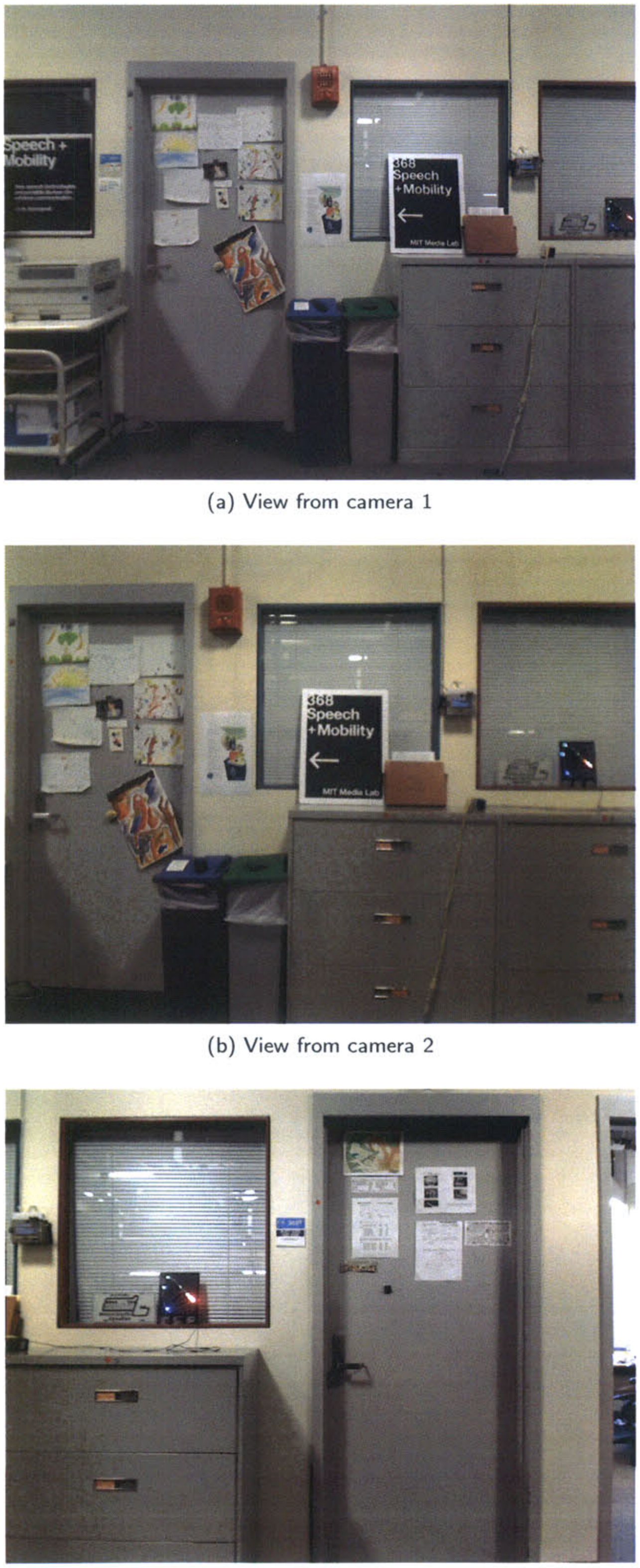
Fig. 4-6 Sample scene: Bulletin Board - original images



(a) View from camera 1

(b) View from camera 2

Fig. 4-7 Sample scene: Bulletin Board - assembled in a 3D image-based model, portrayed from distinct camera viewpoints

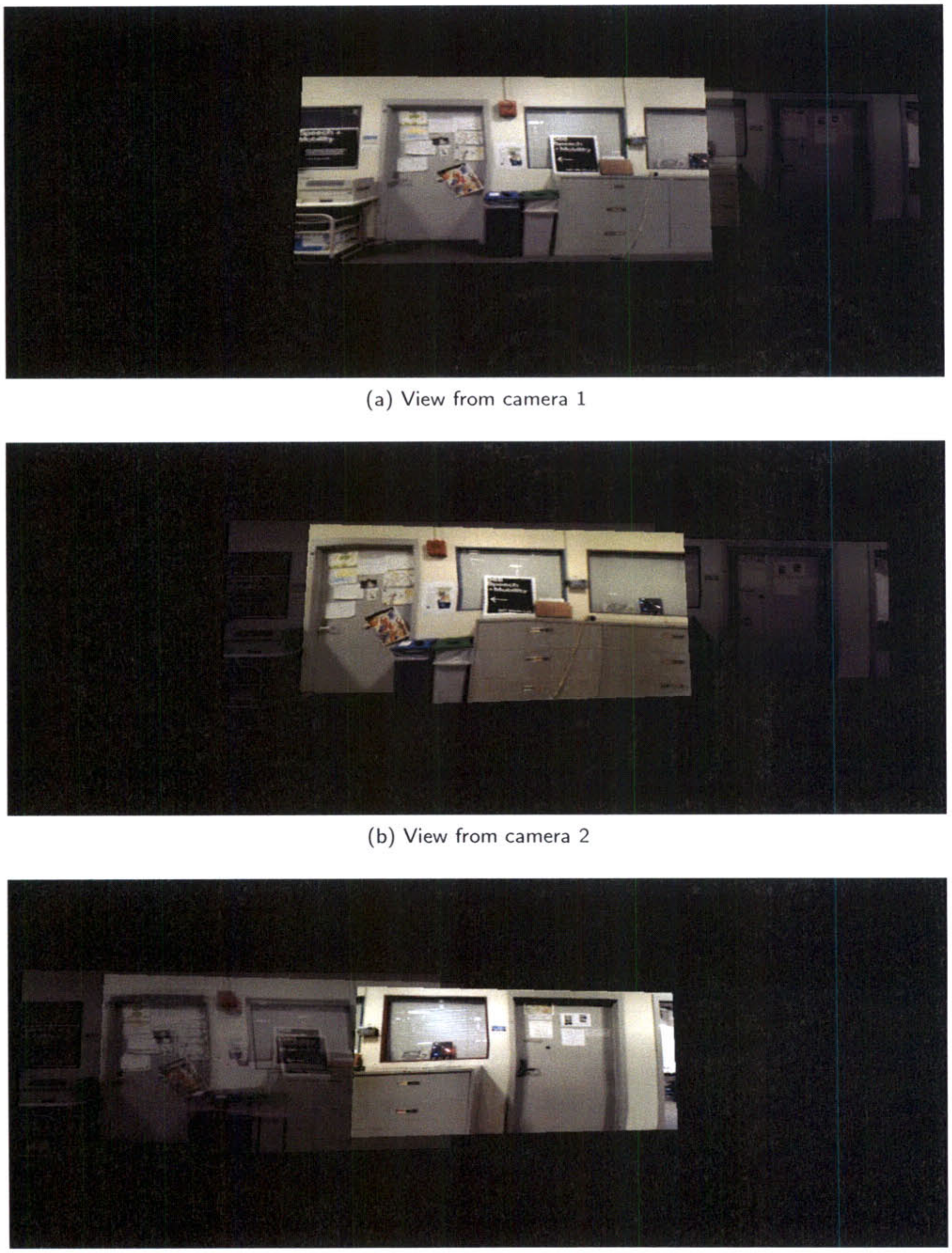


(a) View from camera 1

(b) View from camera 2

(c) View from camera 3

Fig. 4-8 Sample scene: Wall and some furniture close to the wall (camera sensors are parallel to the wall) - original images

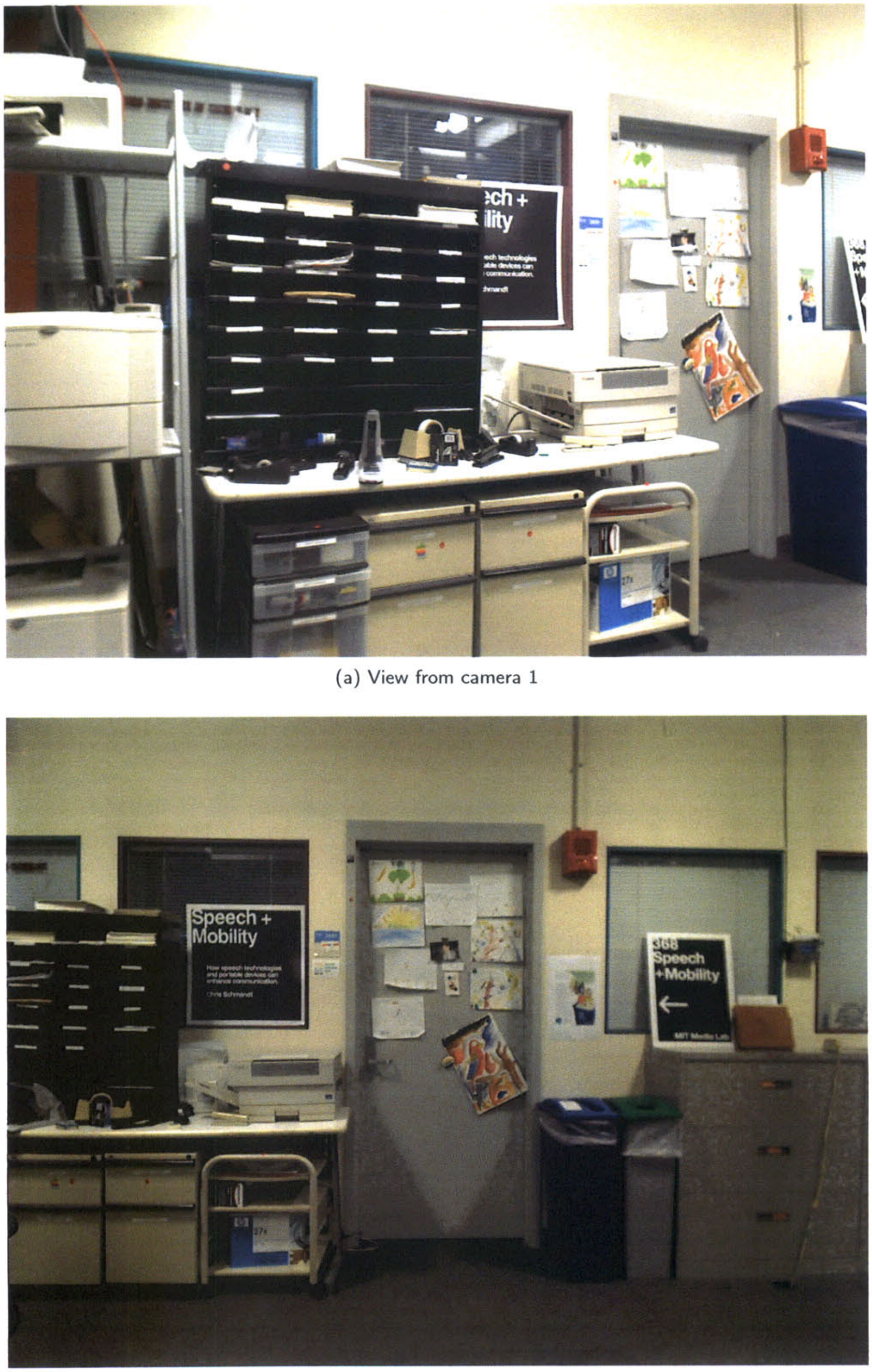


(a) View from camera 1

(b) View from camera 2

(c) View from camera 3

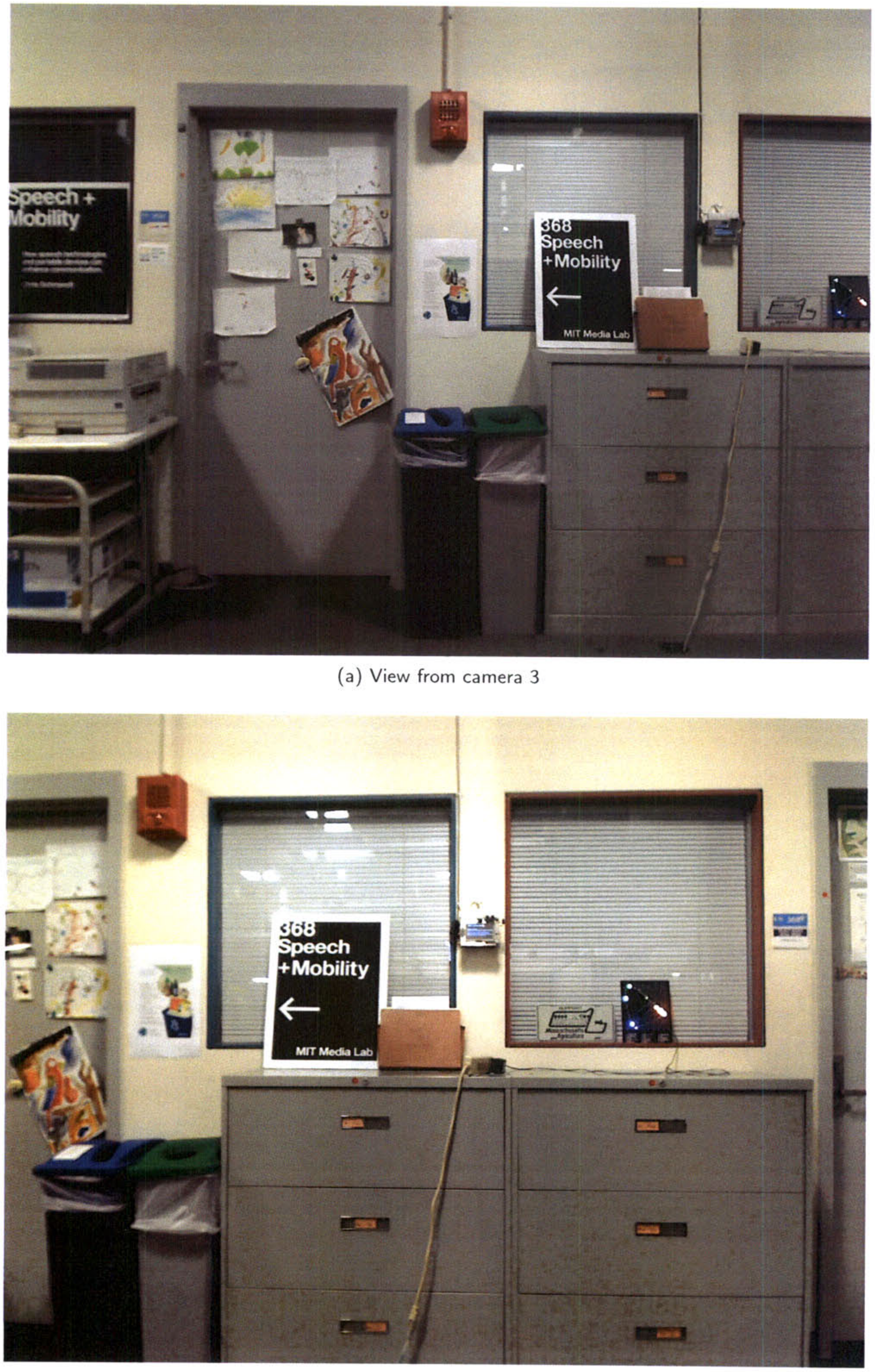
Fig. 4-9 Sample scene: Wall and some furniture close to the wall (camera sensors are parallel to the wall) - assembled in a 3D image-based model portrayed from distinct camera viewpoints


Speech +
Mobility
How speech technologies and portable devices can enhance communication.
Chris Schmandt
368
Speech
+Mobility
MIT Media Lab

(a) View from camera 1

(b) View from camera 2

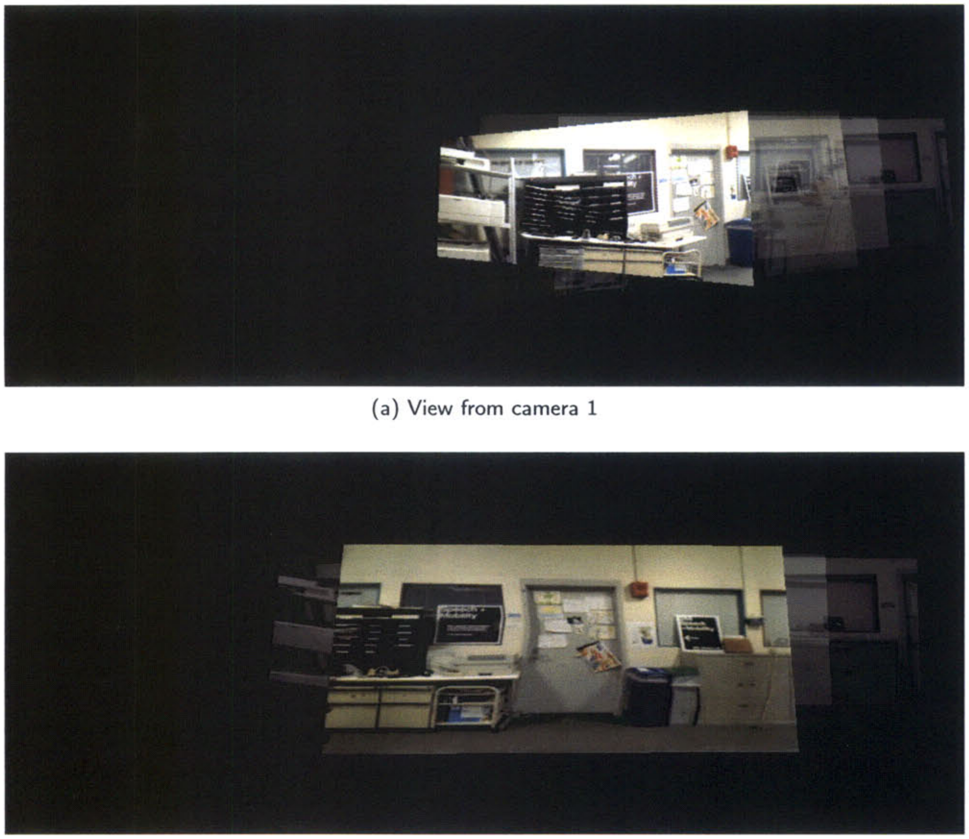
Fig. 4-10 Sample scene: Wall and some furniture close to the wall (not all camera sensors are parallel to the wall) - original images 1 and 2



(a) View from camera 3

(b) View from camera 4

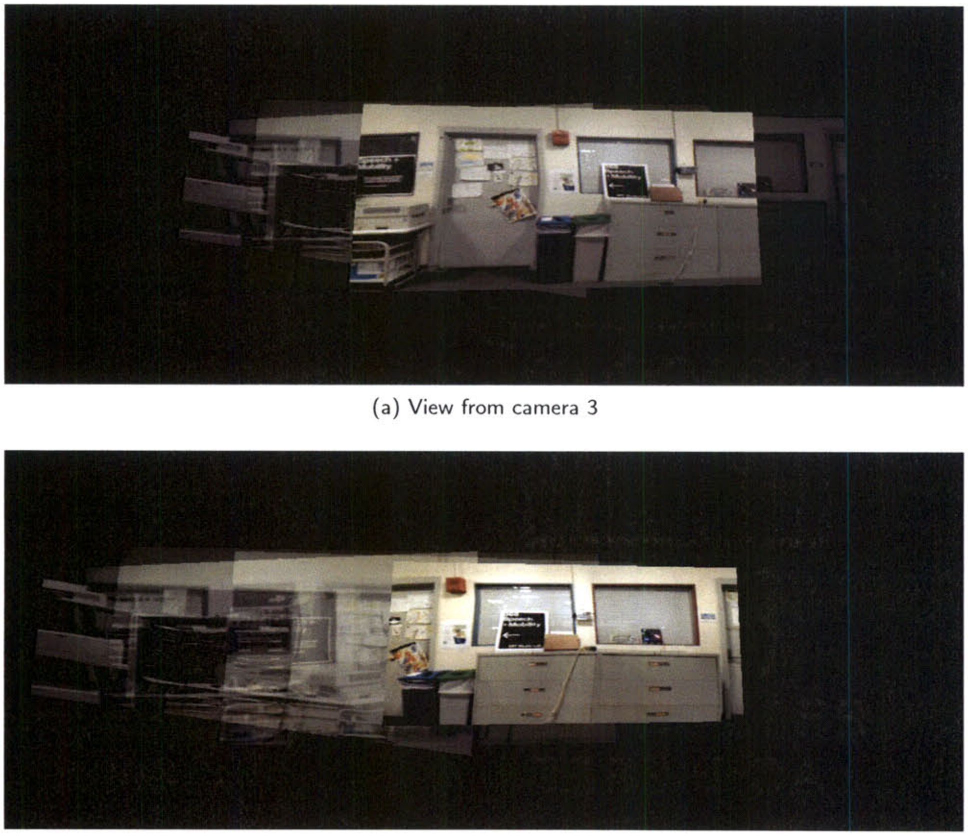
Fig. 4-11 Sample scene: Wall and some furniture close to the wall (not all camera sensors are parallel to the wall) - original images 3 and 4



(a) View from camera 1

(b) View from camera 2

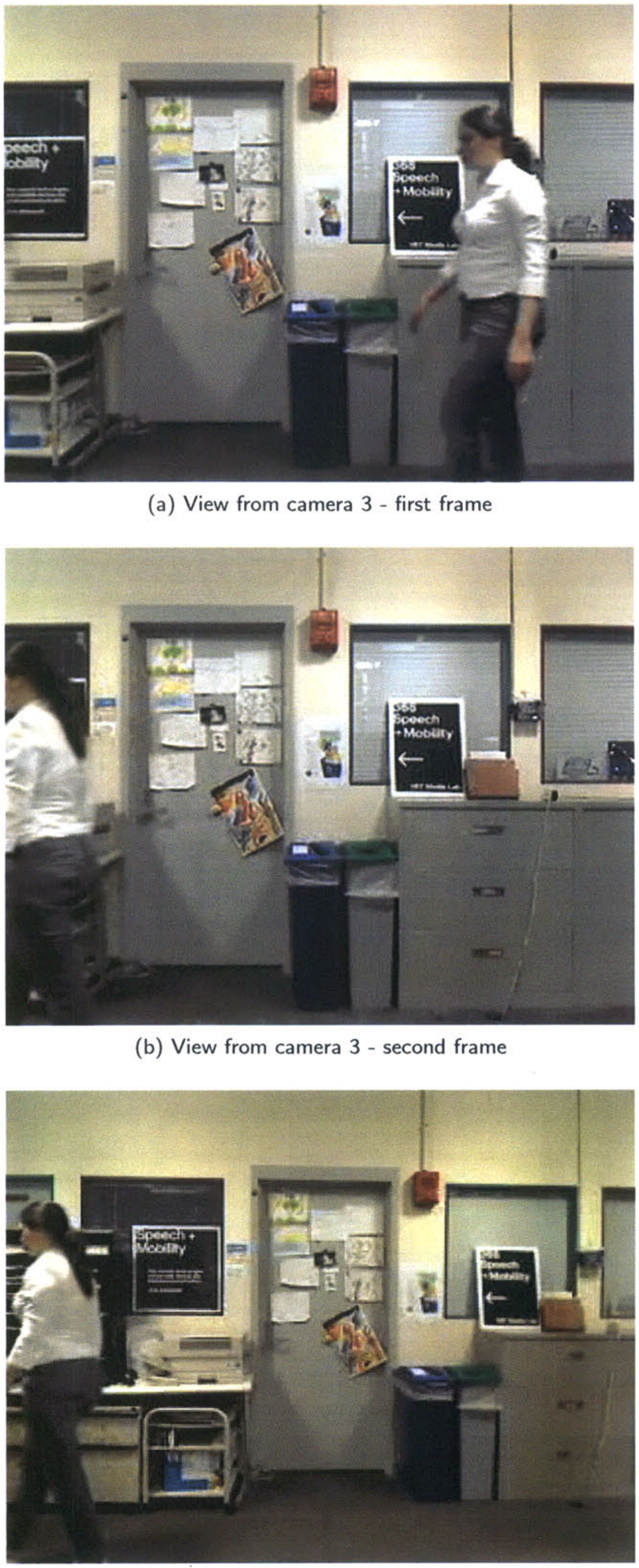
Fig. 4-12 Sample scene: Wall and some furniture close to the wall (not all camera sensors are parallel to the wall) - assembled in a 3D image-based model, portrayed from cameras 3 and 2's viewpoints



(a) View from camera 3

(b) View from camera 4

Fig. 4-13 Sample scene: Wall and some furniture close to the wall (not all camera sensors are parallel to the wall) - assembled in a 3D image-based model, portrayed from camera 3's viewpoint

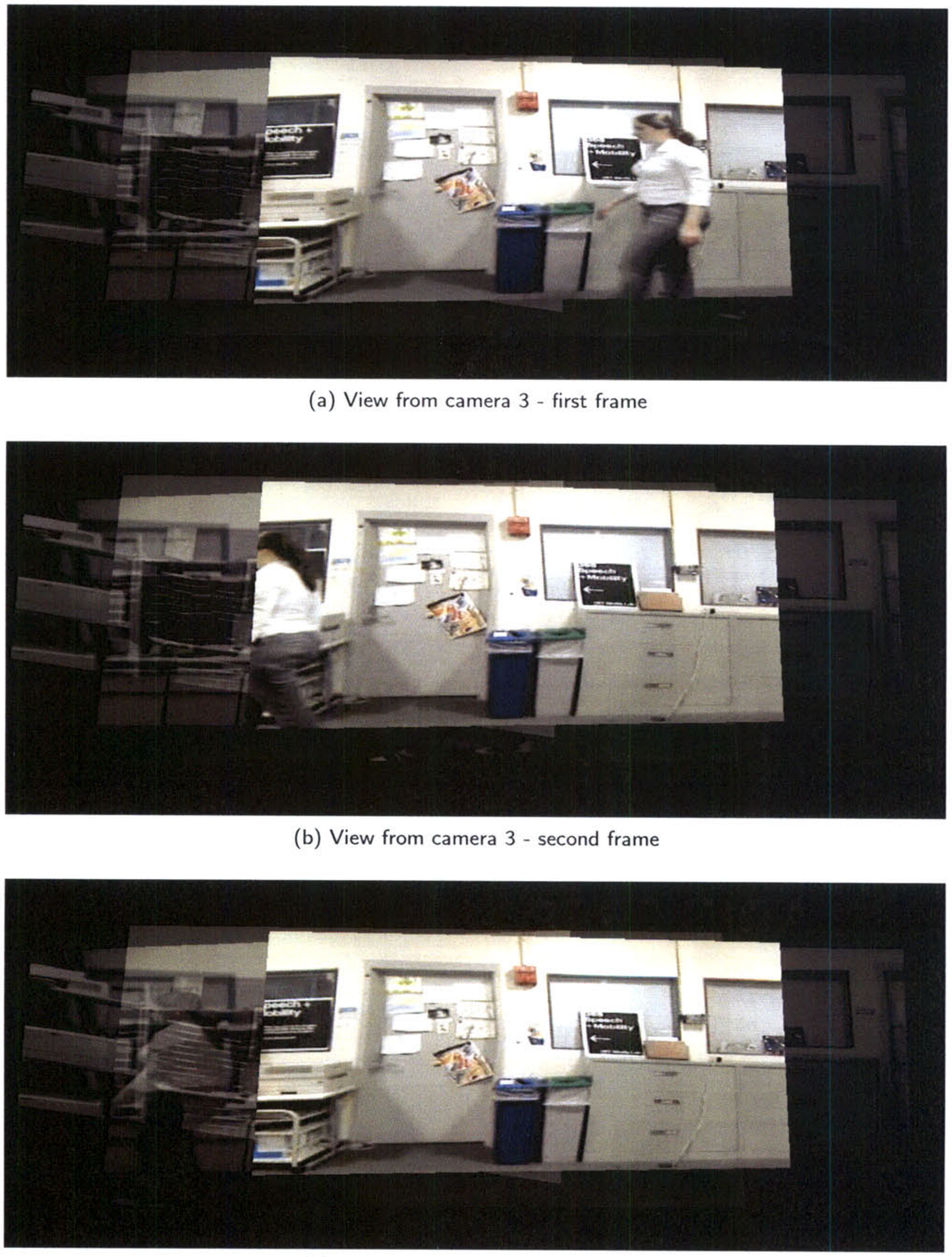


(a) View from camera 3 - first frame

(b) View from camera 3 - second frame

(c) View from camera 2 - third frame

Fig. 4-14 Sample scene: frames of a walkthrough video sequence in front of the Wall and some furniture close to the wall (not all camera sensors are parallel to the wall) - frame sequence from different camera viewpoints



(a) View from camera 3 - first frame

(b) View from camera 3 - second frame

(c) View from camera 3 - third frame

Fig. 4-15 Sample scene: 3D image-based model sample frames of a walkthrough video sequence in front of the Wall and some furniture close to the wall (not all camera sensors are parallel to the wall) - frame sequence from different camera viewpoints

5

Evaluation

In this chapter, we present the process used to evaluate uCom's current prototype. We start by stating the overall purpose of the studies. Next we present our preliminary study, upon which we draw inspiration and learn lessons to conduct the main experiment. Then, our main experimental method is detailed, including its physical setup, research questions, evaluation scenarios and related discussions.

5.1 Purpose of Study

As previously mentioned, uCom's vision is to enhance the feeling of being present in two locations simultaneously. This vision is implemented by enabling real-time awareness of what happens in a remote site by using video assembled to create a visual representation that is coherent with the layout of a remote space. The project's evaluation assesses questions related to the geometric cues portrayed from a remote space. Ultimately, we investigate to what extent uCom improves users' understanding of a remote location's layout.

User tests aim to evaluate uCom's main subsystems: the 3D scene image-based model of the remote space, and the features that allow mapping remote views to local displays. Hence, we are trying to answer two main research questions:

- Does our designed intermediary 3D representation of a space help remote users better understand the space's layout than when using a tiled arrangement?
- Does mapping live views of the remote space to displays in the local space enable a sense that the remote space is integrated into the local physical environment?

The value added from this research is closely aligned with the trend toward ubiquitous video cameras and displays, whether in our homes, work or public environments. The study results are compiled from feedback provided by subjects during the study and from our observations on how users interact with uCom.

5.1.1 Preliminary Study

During the process of designing the user studies, a few participants were invited to give feedback on the proposal. The main purpose was to help us determine the kind of setup and equipment we should use. The idea was to simulate the proposed user study environment in which we portray images of a remote location taken from different viewpoints. Users were asked to make general comments on the setup and also to react to the placement and size of those images.

Each individual was invited to the main location where the final study was expected to be held. Users were four research assistants who work in the same building where this work is being developed. All of them are males, at ages ranging from 24 to 32 years old, and with normal or corrected-to-normal vision. The simulated study setup comprised of a room, with one dedicated chair and table. Easels were placed around the table at different distances from the subject. They were spatially arranged in three concentric semi-circles with three, four and five feet radius respectively, and the user's chair was placed on the semi-circle center point. Figure 5-1 portrays the setup and 5-2 show the diagrams of different positions where the photos were placed during the experiments. The study shows users three different printed photos taken from a single common space, the cafe in MIT building E-15, which is familiar to all users. The sets of images are comprised of photos of different views of the cafe printed in glossy paper in 3 different sizes: 26", 37" and 42" diagonal inches, following 16:9 widescreen standard proportions. Figure 5-3 shows the images used in this study.

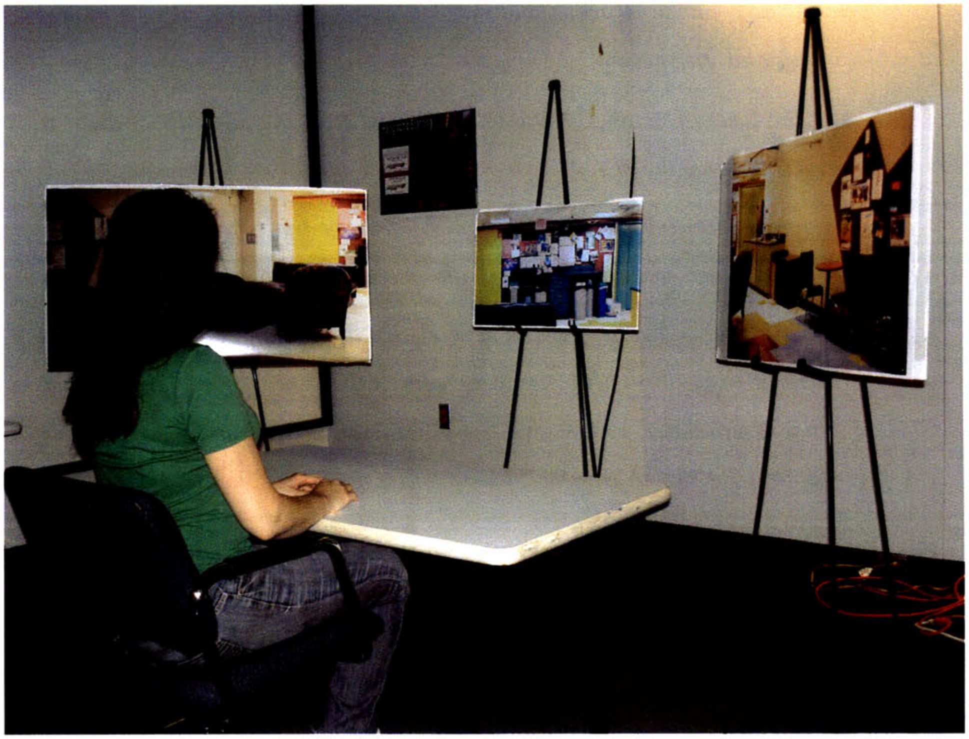


Fig. 5-1 Preliminary study setup

The study protocol starts by introducing the user to the idea of the final system's goal, summarized as "providing visual awareness of a remote space". The user is also informed that photos would be replaced by computer monitors or television sets in the final setup. No further details are provided in order to prevent inducing any bias to user's comments about the experiment. A subject is asked to sit on a chair at a specific position. Then, in each round of the experiment, a set of images portraying three overlapping views of the remote space is placed on the easels around the user. At first, we experiment by placing same-sized images at different distances from the user, at a three, four and five foot radius. Next, images printed on different sizes are presented at the same time.

Throughout the experiment, users are asked to express perceptions, feelings or impressions from seeing different views of the remote space, with different image sizes, at different distances. A sample of user comments are listed below.

1. Quotes on the size of displays, and display to user distance:
    - *"A large image is overwhelming, but more immersive, attention grabbing and distracting."*
    - *"The level of detail I'd like to see depends on the situation I want to see. It's case by case."*
    - *"Image size depends on the application. If I am trying to establish a conversation via the screen, the larger it is, the better."*
    - *"My distance to the screens depends on the social context: the activity I want to perform with them, or some other activity I might be focused that is unrelated to the screens. For instance, if I am working on my laptop, I'd like to be very focused to it, not to be distracted."*
2. Users' comments on the arrangement of the displays:
    - *"If the images have a good default placement, they are like furniture as they don't need to be moved very often."*
    - *"Its important to have the screens arranged coherently. Otherwise it's hard to create a mental model of the other space."*
    - *"If you optimize the placement of the screen, and derive a sense out of this arrangement, you can naturally keep an eye on them all the time."*

The conclusions drawn from users' comments are summarized as follows:

- The optimal size of the displays and the distance between displays and users are situation specific. First, they relate to how much attention a user wants to pay to what happens at a remote space. Second, it depends on how distracting those images are to other activities a user might be performing.

- The arrangement of the displays should be coherent to how the images align in space. They should match the arrangement of the remote space they portray.

5.2 Main Study

5.2.1 General description

Based upon the preliminary study, the research methodology for the main experiment involves asking a subject to interact with our system in order to perform a set of predetermined tasks. The investigator remains in the room at all times, observing while the user interacts with the system, asking questions and taking notes on the users reasoning process and comments. This process is not exactly structured as an interview, but it opens and closes with questions about the experiment. Photos of the experiments are taken to keep a record of the equipment placement proposed by the subjects. None of the subjects appear in those photos.

The study protocol starts with one subject at a time being invited into a room, with its table, chair and multiple displays installed on rolling floor stands. The subject is asked to sit down and is briefed about how to interact with the system by using a keyboard and remote control, and by also slightly repositioning the displays. In the experimental setup live video feeds of a remote space are replaced by still images. This decision was due to practical reasons, as it creates a more controlled environment. First, still images can portray remote locations without requiring video cameras to remain in place throughout the whole experiment. Live video could lead to privacy concerns among individuals not participating in the study, as the portrayed area is regularly used by multiple people. Second, images can convey the layout of the remote space.

5.2.2 Physical setup

The study is physically setup in rooms of MIT building E-15. One of the rooms contains several fixed displays mounted on rolling floor stands. Video cameras were temporarily installed in another room of the building. During the experiments any of the pre-recorded camera views could be connected to any of the monitors in the experiment's room. The 3D intermediary representation of one space could also be viewed from any monitors at the other space. No video recording takes place during the experiment.

It's worth mentioning that the locations of the user studies are representative of a work environment: a conference room and common work areas, all situated in an office building. This setup was chosen to reflect a work environment. But it's worth mentioning that uCom can be used to connect any two bounded locations, as long as multiple cameras are available in one site and displays are available on the other site. Some system restrictions were detailed in chapters 2 and 3. They comprise mostly the placement of the cameras to provide a minimum overlap between "neighbor images" and minimum image resolution.

To evaluate the system, users were assigned tasks related to setting up how the remote space should be mapped to the local space. The quality of the visual mapping between the spaces will naturally be affected by the equipment restrictions: quantity, position, resolution, and so forth of both cameras and displays, besides the screen size and mobility of the displays.

The list of equipment utilized in the study follow below.

- Five LCD display equipment: two with 32" and one with 42", one 19" and one 13" computer monitors;
- Three rolling floor stands in which the larger displays were mounted;
- Four webcams, capable of acquiring still images with a 1024x768 pixels minimum resolution;

- Five computers.

It's worth mentioning that the display sizes were chosen for three main reasons: (1) the feedback given by the participants of our preliminary assessment, (2) the considerable availability of video display equipment with those sizes, and (3) the size of the room with respect to recommendations on television viewing distance. The Society of Motion Pictures and Television Engineers (SMPTE) [1] recommends that the screen size for a home theater use should occupy a 30° field of view - in the horizontal plane. This corresponds to a viewing distance that is about 1.9 times the screen width. The two kinds of displays we used were televisions with 19", 32" and 42" diagonal, with 16", 28" and 37" width respectively. Therefore, the viewing distance should be approximately 30", 53" and 70". Those distances were appropriate to the room where the experiments took place.

5.2.3 Subjects

We recruited eight individuals with ages ranging from 18 to 65 years old. The criteria for inclusion or exclusion in our study were:

- subjects were required to have normal or corrected-to-normal vision;
- subjects were required to have physical ability to slightly reposition displays placed on rolling floor stands, which is not physically stressful;

Additionally, subjects were asked if they were familiar with the layout of the location portrayed in the images shown during the experiment. It should be noted that this was not used to exclude subjects from the experiment. Yet we only describe the results of the eight users who were not familiar with the portrayed location, in order to avoid any bias in the responses due to previous knowledge of the remote space layout.

[1] SMPTE Recommended Practice 166-1995: http://www.smpte.org

The user tests were open to volunteers from the MIT community. Study participants were recruited using email lists and posters spread throughout the campus. Subject were offered $10/hour compensation in gift certificates for the time and effort associated with participating in the study.

5.2.4 Formal Approval

This study has been reviewed and approved by the MIT Committee On the Use of Humans as Experimental Subjects (COUHES), with protocol number 0904003205. The major elements of the application are described below:

- Photos of the experiments are taken to keep record the equipment placement and drawings proposed by the subjects. None of the subjects appear in those photos.
- There is no audio or videotaping.
- All study data is coded, so it's not associated with any personal identifier.
- No information about the research purpose and design will be withheld from subjects.
- Study data is stored and secured in the MIT Media Laboratory servers without any user identification.
- The study takes less than 1 hour per each subject.

5.2.5 Tasks

The subject is asked to perform the following tasks during the experiments: (1) navigating through the system's graphical user interface with a remote control and keyboard, (2) repositioning monitors placed on rolling floor stands, and (3) answering questions asked by the investigator. The research questions about which we draw conclusions, and the respective questions we ask the subjects follow below.

Research question:

Does our designed intermediary 3D representation of a space help remote users better understand the space's layout than when using a tiled arrangement?

Experimental steps:

1. User is shown the intermediary 3D representation of video views from the remote space.
   (a) Use the keyboard to navigate through the 3D representation of the remote space.
   (b) Explain your perception of the relative positions of the main furniture, doors and windows of the remote space. Describe it verbally and with gestures, or sketch it on paper.
2. User is shown a tiled arrangement of video views from the remote space, which resembles a surveillance system display.
   (a) Use the keyboard or mouse to navigate through the different views of the remote space.
   (b) Explain your perception of the relative positions of the main furniture, doors and windows of the remote space. Describe it verbally and with gestures, or sketch it on paper.
3. User is asked questions about both representations.
   (a) Which representation helped you better understand the layout of the remote space? Was there a relevant difference between them in this case?
   (b) Are the available views enough to understand how the remote space is arranged?
   (c) Would you prefer to have more camera views? If so, how many more cameras?

(d) Would you suggest any different viewpoint that is not currently available?

(e) Can you point out any view that was not particularly satisfying and why?

Note: For each subject, users are first shown the 3D image-based model and then the tiled view. This decision is based on the fact that the 3D image-based model is not a natural representation to which users are accustomed.

Research question:

Does mapping live views of the remote space to displays in the local space enable a sense that the remote space is integrated into the local physical environment?

Experimental steps:

1. User is told that the focus of the experiment changed to using the displays available in the room to have a sense of the geometry of the remote space. New tasks will be requested.

   (a) Considering the position of the displays as fixed, experiment switching the view shown in each display.

   (b) From now on, you are allowed to freely reposition the displays. So, place the displays as close to you as still feels comfortable.

   (c) Now, place the displays as far from you as you can still see them clearly.

2. User is asked questions about the tasks that have just been performed.

   (a) How did you decide about which view should be shown in each display?

   (b) Did you feel immersed in the remote space for having those views constantly available around you? Please comment.

   (c) Which factor do you think is more important on positioning the views: the geometric position between images or the remote situation you are trying to follow?

(d) Do you feel distracted by having those remote views around you?

(e) Discuss about the multiple views of the remote workspace already mapped to the local space, drawing comments on whether they can convey a sense of how the remote space is arranged.

(f) Were you satisfied with the size of the displays? Were they too big or too small for this specific room size?

(g) What did you think of repositioning the displays? Was it useful in this case? If not, why?

Post-experiment questionnaire:

- In what kinds of situations would you be interested in using this system? Connecting you with which other people? Or between which locations?
- Discuss the usability of the interface.
  - Was it easy to navigate the sets of images?
  - Were the mouse and keyboard controls easy to understand and operate?
  - Were there any limitations?
  - Would you suggest any modifications on how to interact with the system?

5.3 Results and Discussion

The results of the user studies were obtained from the eight users who were not familiar with the scene, i.e., they had never been to the portrayed location. The images used in the experiment show a corner of a room and two adjacent walls, which can be seen in Figures 5-4. The physical setup of the room in which subjects performed the experiment is showed in Figure 5-5. The 3D image-based model is shown from an estimate of images 1 to 4's viewpoints in Figure ??, and the tiled arrangement of the views is shown in Figure 5-10.

Next we present and briefly discuss the major results from the study, from which we attempt to draw conclusions about the two main research questions: (1) "Does our designed intermediary 3D representation of a space help remote users better understand the space's layout than when using a tiled arrangement?", and (2) "Does mapping live views of the remote space to displays in the local space enable a sense that the remote space is integrated into the local physical environment?".

The answers and comments obtained from the subjects are grouped in related topics, presented as follows:

- Reasoning process towards understanding and explaining the layout of the remote space.
- Reactions to the two representations of the space: the 3D image-based model and the tiled view.
- Available views' ability to enable subjects' understanding of the remote space's layout.
- Reasoning process behind choosing which image to be shown in each display.
- Reactions for having remote views constantly available around subjects: immersion, distraction and comfort levels.
- Impressions on the ability to reposition the displays.
- Potential applications foreseen to the system.
- Feedback on the user interface.

Reasoning process towards understanding and explaining the layout of the remote space

Most subjects preferred to verbally explain their perceived mental model of the space while using gestures to refer to the images and their relative positions. Only two out of the eight subjects decided to sketch the space. Yet only one of them could actually draw an accurate model of the space. This pattern suggests that it's not so easy to express one's understanding of a 3D space's layout. The most common reasoning process to create a mental model of the space's layout was to use the objects portrayed in the images as references. Figures 5-6 and 5-7 show edge-filtered versions of the images used in the study, which are used to highlight and label the objects used as references. Only one out of eight users tried to connect the images by matching the image edges - the frames of doors and windows. These attempts were particularly unsuccessful with the 3D image-based model, as the system introduces distortions on edges. Additionally, some users made common mistaken assumptions about the scene. First, a couple of subjects initially thought that the four images represented four walls of one closed room. Second, one user assumed that the same object appearing in more than one image actually meant that there was more than one instance of the object in the scene.

Reactions to the two representations of the space: the 3D image-based model and the tiled view

Concerning user's reaction to the 3D image-based model shown in Figure ??, almost all users were able to find out how the space was laid out by matching objects present in multiple images. Most subjects remarked that the horizontal arrangement of the images in the 3D image-based model and the navigation controls made it easy to understand how they related. A couple of users faced difficulty in understanding the space layout using the 3D image-based model. Yet only one of them was unable to explain the space's layout properly. Even after being told that the 3D image-based model introduced distortions in the images, two out of eight users had particular difficulty in connecting the images as they focused too much on the fact that some edges were distorted and were also confused with the blending or transparencies on overlapped images. Yet, even when shown the tiled view, those subjects still took longer than others to explain how the space was laid out. This fact might suggest that those particular users have a harder time in perceiving a 3D space. A particular reaction was noteworthy among those subjects: they had some difficulty in perceiving that image 2 in Figure 5-8b portrayed a 90° corner between two adjacent walls. It should be noted that the corner of the scene portrayed in this image was essential to users' understanding of the spatial layout.

Concerning user's reaction to the tiled arrangement shown in Figure 5-10, it was straightforward to most users. Most of them used it to confirm what they had thought of the space's layout from the 3D image-based model. The same two users who had difficulties in understanding the space from the 3D image-based model still had a hard time with the tiled view. Yet, it was a bit easier for them as the tiled view didn't distort the images.

Comparing user's reactions to the two representations of the space, most subjects seemed to prefer the tiled view. Most of them remarked that the distortions in the 3D image-based model made it harder to understand. Only two out of eight subjects claimed that the 3D model easily enabled them to create a mental model of the remote space, particularly due to the right to left navigation controls that followed the arrangement of the images and the position of its cameras in 3D space. For those two particular users, the fact that the tiled view presented the images in a random order did not overcome the distortions of the 3D image-based model.

Available views' ability to enable subjects' understanding of the remote space's layout

All users claimed the available views were enough to enable them to create a mental model of the remote space. Two out of eight users particularly suggested that the corner of the room, shown only in image 2 5-4b should be portrayed in other images from different viewpoints to make it easier for them to understand that it represented a 90° angle between two adjacent walls. As previously mentioned, two out of eight subjects thought at first that the four images represented four walls of a closed room. In addition, all subjects referred to image 3 (Figure 5-4c) or image 4 (Figure 5-4d) as not being necessary to aid them into understanding the layout of the remote space. It is understandable, as both images portrayed almost the same objects. Yet image 3 shows less objects as it is a narrow angle version of image 4. Interestingly, subjects would usually refer to the tiled view to point to the image that was not particularly helpful to create a mental model of the geometry of the space.

Reasoning process behind choosing which image to be shown in each

display

Most users tried to enable the images in the monitors so as to match the geometric position of the portrayed scene. Most subjects didn't use the monitor that is placed on the desk. The reason could be that all 3 most important images that help building the 3D mental model of the space could already be seen in the three larger screens placed in floor stands. Additionally, the images not usually shown on the displays are either image 3 (Figure 5-4c) or image 4 (Figure 5-4d), which portray almost the same objects. A couple of users showed the same image in all screens. The reason was not so clear. But as we purposely didn't provide guidelines on the expected motivation behind displaying remote images, some users made unexpected decisions, such as: showing the images that seemed farther away, or different images that seemed to have been captured at a similar distance from the wall (narrow or wide angle) or images that had similar brightness levels.

Reactions for having remote views constantly available around subjects: immersion, distraction and comfort levels

Most subjects' comments regarding immersion state they don't feel immersed. But they have a "sense of being able to monitor the remote space" or "it doesn't feel immersive, but only portrays a logical geometric correspondence". A user remarked that when displays are closer to each other, it feels more immersive. Concerning brightness of the screens and their potential distraction, half of the subjects claimed the brightness from the screens was distracting, while the other half said it was not distracting at all. But we should take into consideration that the part of the study that had the monitors on lasted 20 minutes at most. Having them on for a whole day would possibly generate different reactions. A few users (two out of eight) suggested that the screens should be placed as far from them as the images could still be seen properly, so as to prevent the brightness of the monitors to be cumbersome. Half of the subjects claimed the monitors were too close to them, while the other half felt comfortable about their current distance with respect to the screen size. A few users (two out of eight) made comments about the fact that different sized displays and the overlap between images made it awkward to have them closer and still feel like the images were connected seamlessly.

Impressions on the ability to reposition the displays

Most subjects (six out of eight) took advantage of the possibility of moving the displays. They made positive remarks about the flexibility the system provides, allowing displays to be moved to fit users' needs and changes to the furniture layout. Among the main common remarks, we have the following. First, A few of the users (two out of eight) moved the screens away - seen in Figure 5-12, as the screen's brightness felt disturbing to them. Second, some users placed the screens close together, with the borders of the monitor touching each other, and the image showing the corner of the scene at a 45° angle between the other two perpendicular images - seen in Figure 5-11. They claimed the intent to create a kind of panel with the multiple screens portraying the remote space, yet the border of the equipment felt like a barrier between the images. Third, for other users (two out of eight) it made sense to push the display showing the image taken at a wider angle farther away. So, they pushed away only the monitor that showed the corner of the scene.

Potential applications foreseen to the system

Users were suggested the possibilities of using the system in both a work environment and a home application. Most proposed its use for a surveillance application: in-store security cameras, home outdoor surveillance, monitoring children or pets. Others were too focused in comparing it to end-user videoconferencing like skype or corporate videoconference rooms. One particular user couldn't foresee any potential applications for the system. Most of them reacted positively to our teaser proposal of a "videoconference system that is always on". But some made comments about possible distraction to one's main task. Yet one user made comments about privacy and the arrangement of displays, such as: if monitoring a personal-related situation (such as kids or pets), one would prefer to have them shown in smaller screens, perhaps a secondary monitor on the desk, in order to prevent from sharing private moments with passers-by. Additionally, it was clear to us that users emphasized the need to maintain a geometric correspondence of the physical screens and the remote space. Some users even stated that the arrangement of displays wouldn't change if they were interested in following on a specific view. Some users made remarks in practical terms, such as: the 3D model is more economical, but it doesn't allow all views to be seen at once. On the contrary, the use of multiple monitors and the tiled arrangement allow all images to be seen at once, while the large size of the screens make it even easier to see all of them at once. Another user remarked that using multiple screens was a too expensive and cumbersome setup. One single large screen showing the 3D image-based model would be more efficient cost-wise, and would make the room less cluttered.

Feedback on the user interface

Most subjects pointed out that the remote control was a fairly easy to use and logical interface to switch the views shown in each display. Some suggested they should be able to point the remote directly to the selected screen instead of to one single computer that controls all the screens used in the experiment. This is a system limitation we would certainly improve in a future development.

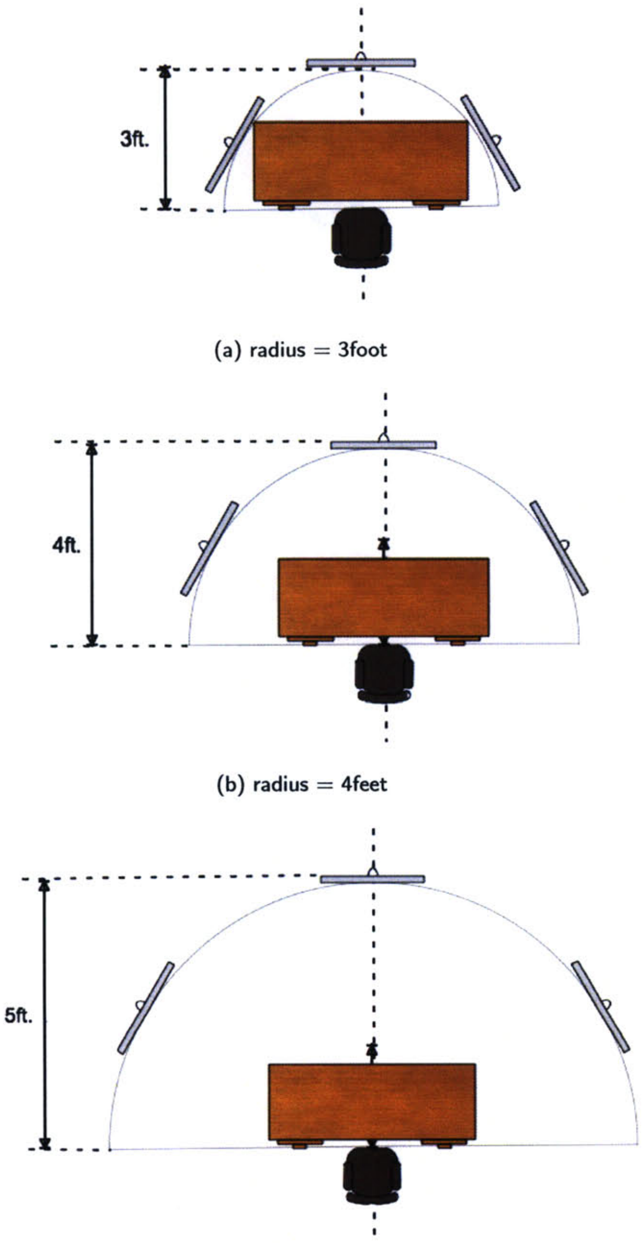
Concerning using the keyboard to navigate the 3D image-based mode of the scene, most subjects solely used the key that switches the selected camera's viewpoint. The use of arrows did not add much to the experience. We perceived a need to switch between cameras with respect to their relative position in multiple axes. In this use case, the scene was composed of images that were laid out horizontally. Switching cameras from right to left was intuitive. But in the case of images with camera viewpoints that don't necessarily align well horizontally, we should better allow users to switch between cameras using at least two sets of axes: x (left to right and right to left), and y (up to down and down to up). Yet, most users stated that the keyboard functions were easy to use. One particular user mentioned that it was cumbersome to memorize the functions of each key, as they were not intuitive. Some suggested a few modifications, such as: (1) the ability to switch cameras in both left to right and right to left direction, (2) the use of the arrow keys to switch between cameras, (3) the use of a mouse to switch between views.

5.4 Experimental Conclusion

The experiments helped us draw conclusions about our two main research questions: (1) “Does our designed intermediary 3D representation of a space help remote users better understand the space’s layout than when using a tiled arrangement?”, and (2) “Does mapping live views of the remote space to displays in the local space enable a sense that the remote space is integrated into the local physical environment?”.

Comparing user’s reactions to the two display formats representing a remote space, most subjects seemed to prefer the tiled arrangement rather than the 3D image-based model. Yet, most of them could effectively create a mental model of the remote space using the 3D image-based model, particularly due to the right to left navigation controls that followed the arrangement of the images and the position of its cameras in 3D space. The major disadvantage of the 3D image-based model is the distortions it introduces on the images, making it harder for users to identify objects in the scene to be used as references for a mental model of the space’s layout.

Concerning mapping views of the remote space to displays in the local environment, our experiments suggest a potential application for constant visual awareness, but not necessarily immersion in the remote space. In fact, most subjects suggest use cases related to remote surveillance and monitoring. Additionally, the experiments indicate that most users try to enable the images in the local monitors so as to match the geometric position of the portrayed remote scene. They also made positive remarks on the flexibility provided by moving the displays, in order to (1) better match the layout of the remote scene, (2) to potentially fit users’ changing needs, and (3) to prevent the brightness of the screens from distracting users from other activities.

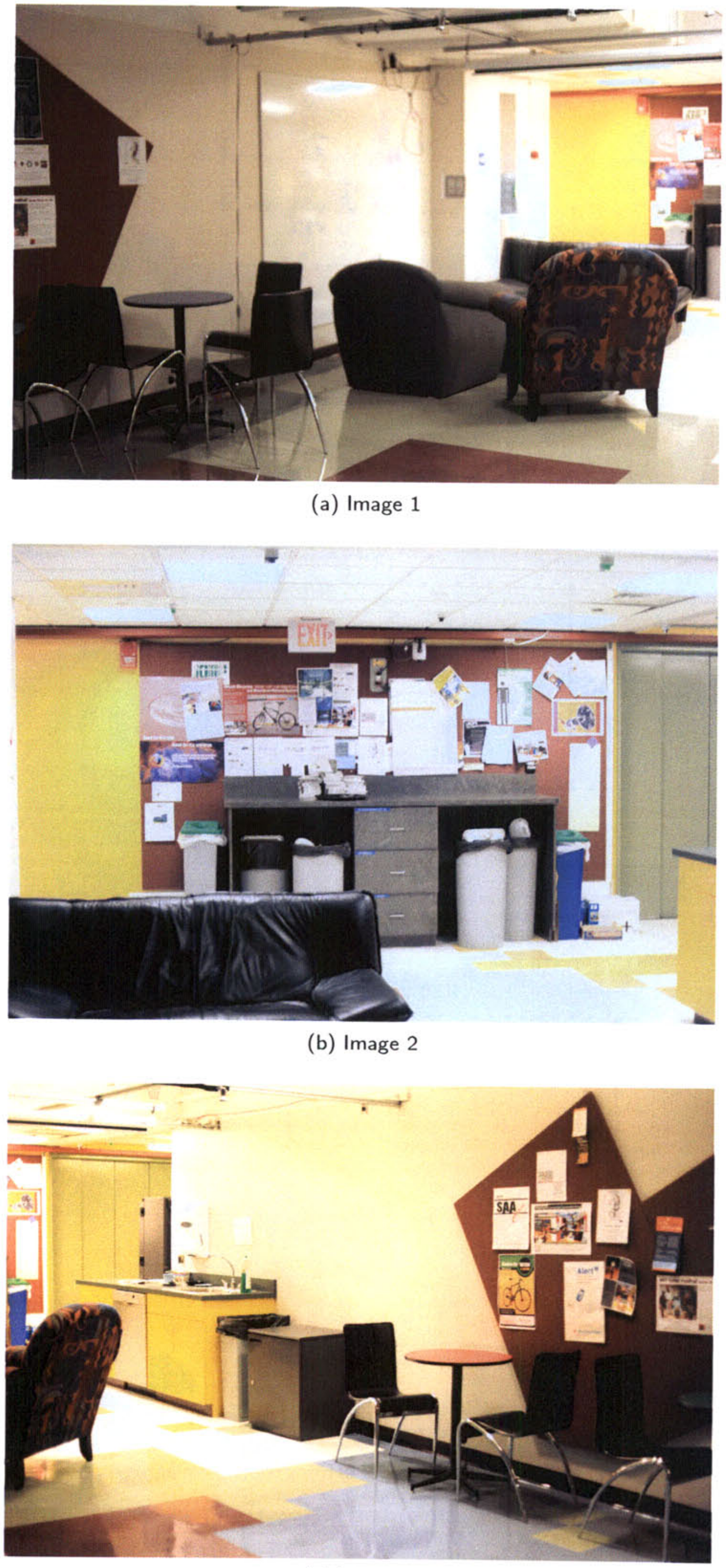


(a) radius = 3foot

(b) radius = 4feet

(c) radius = 5feet

Fig. 5-2 Preliminary study setup - top-view diagrams

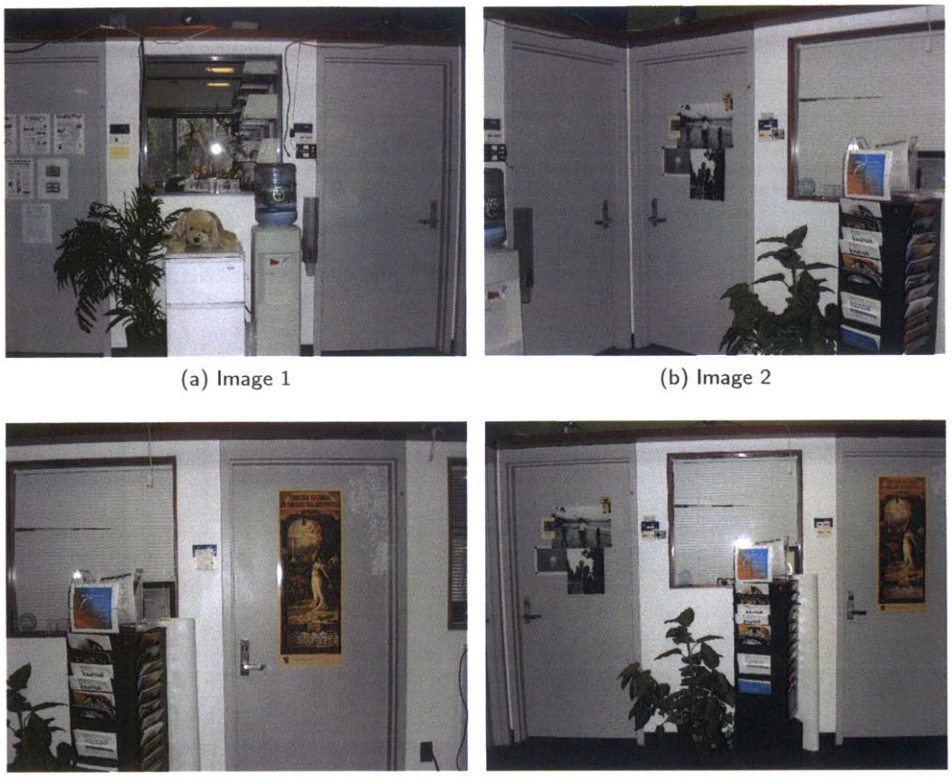


(a) Image 1

(b) Image 2

(c) Image 3

Fig. 5-3 Images used in the preliminary study



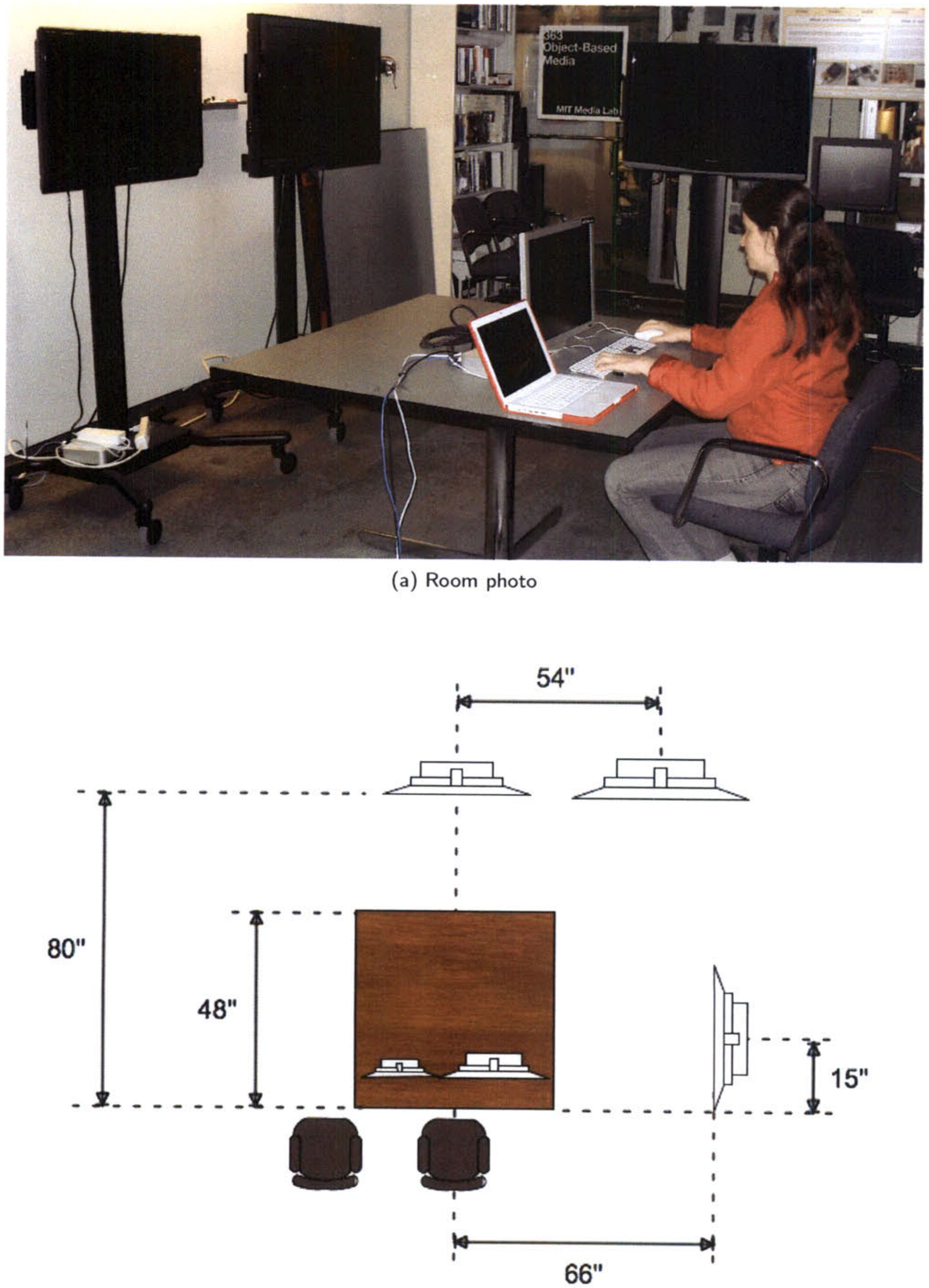
(a) Image 1

(b) Image 2

(c) Image 3

(d) Image 4

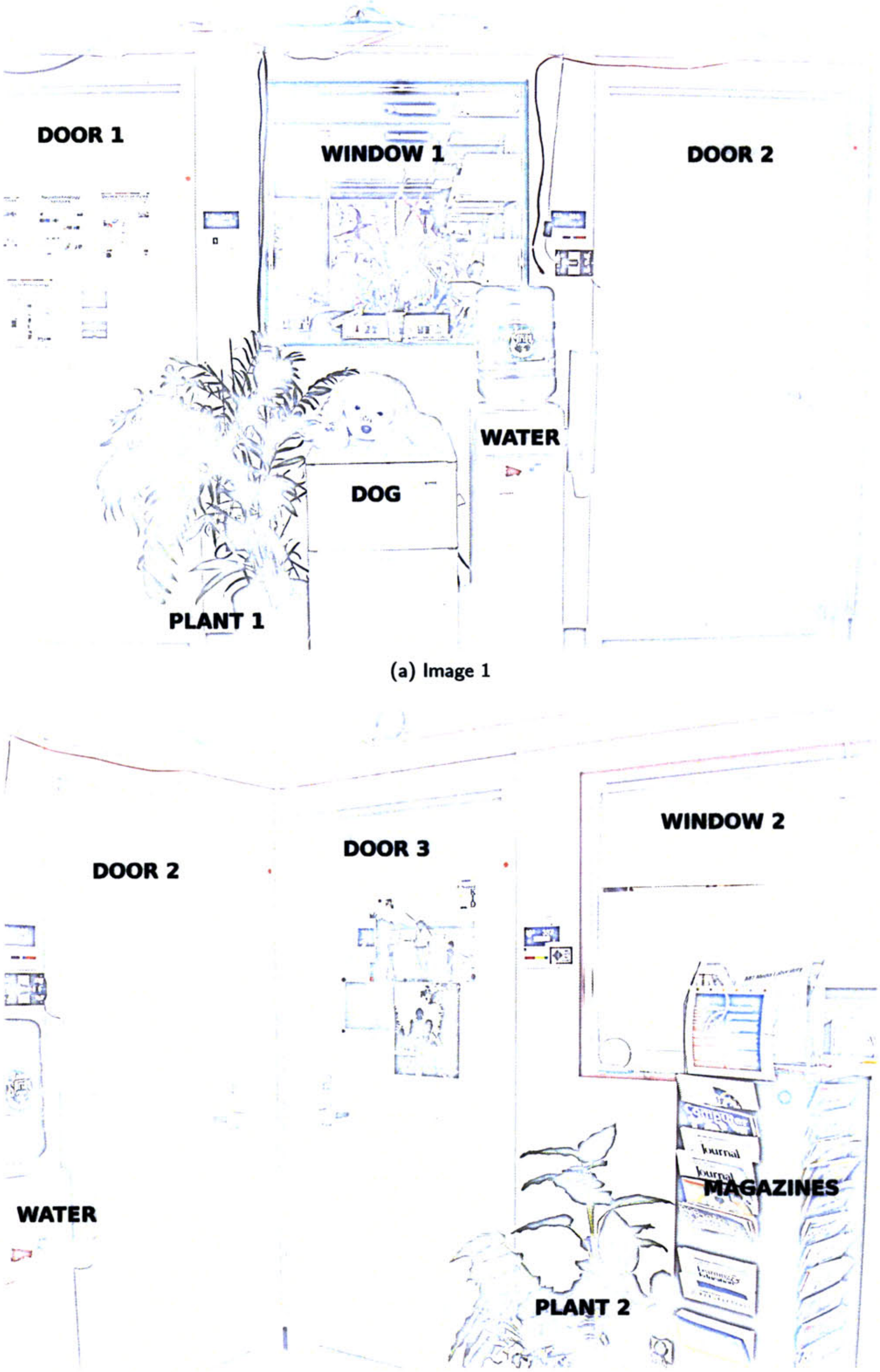
Fig. 5-4 Original images used in the experiment



(a) Room photo

(b) Top-view diagram

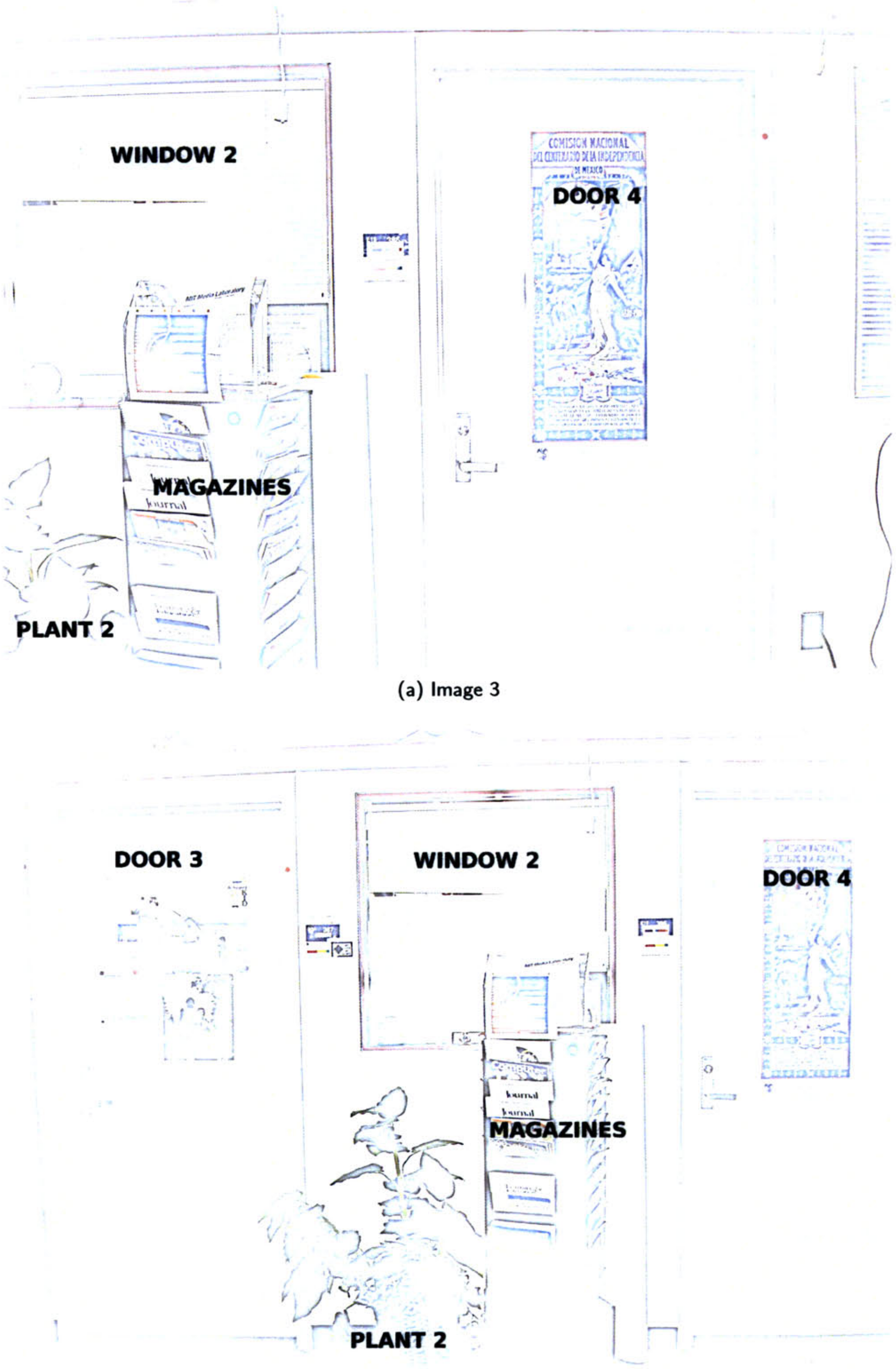
Fig. 5-5 Experimental setup - original setup before the experiment starts



(a) Image 1

(b) Image 2

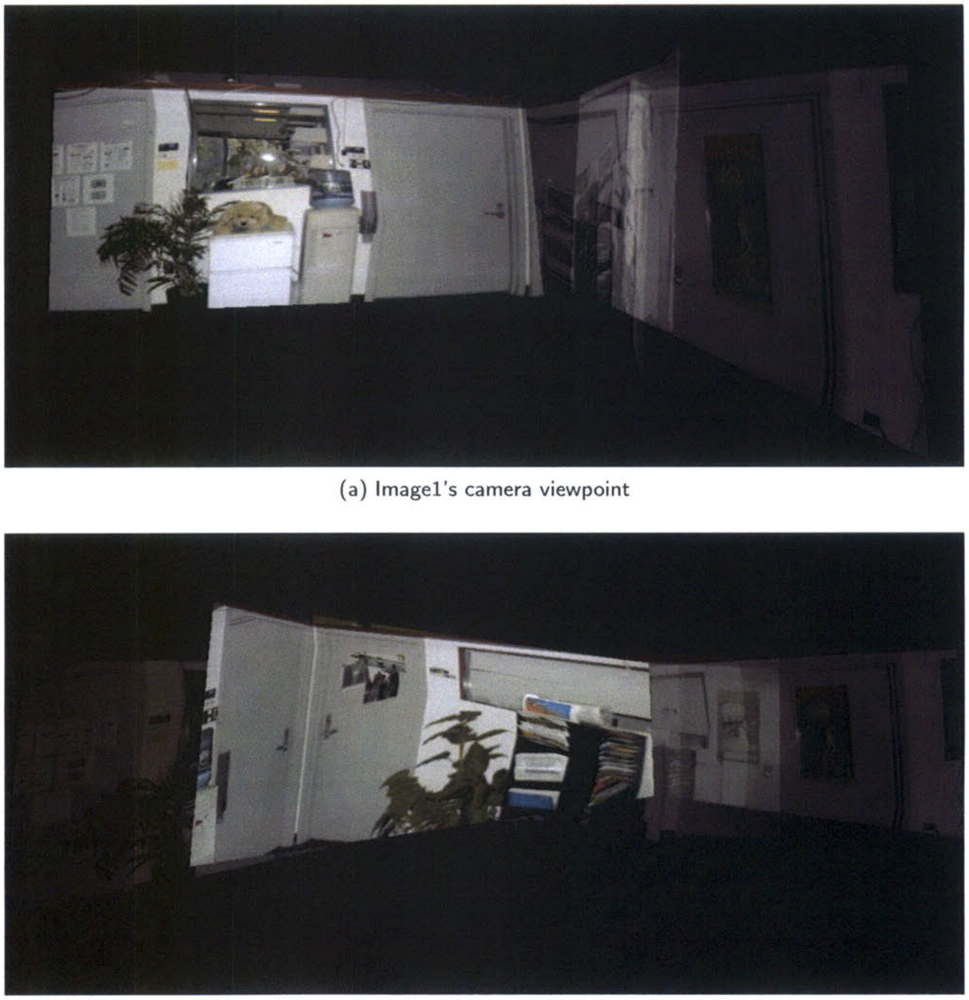
Fig. 5-6 Filtered images 1 and 2 used in the experiment - objects used as references are highlighted and labeled



(a) Image 3

(b) Image 4

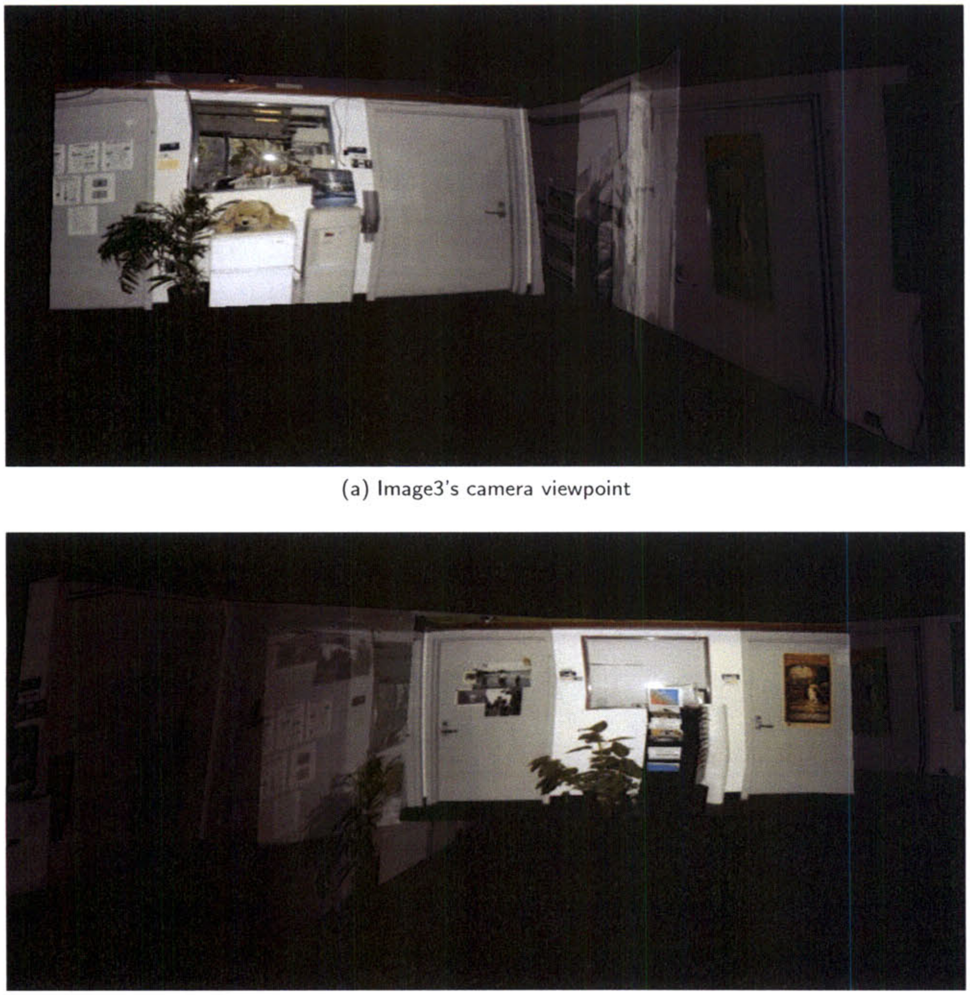
Fig. 5-7 Filtered images 3 and 4 used in the experiment - objects used as references are highlighted and labeled



(a) Image1's camera viewpoint

(b) Image2's camera viewpoint

Fig. 5-8 Experiment - 3D image-based model viewed from the viewpoints of cameras 1 and 2



(a) Image3's camera viewpoint

(b) Image4's camera viewpoint

Fig. 5-9 Experiment - 3D image-based model viewed from the viewpoints of cameras 3 and 4

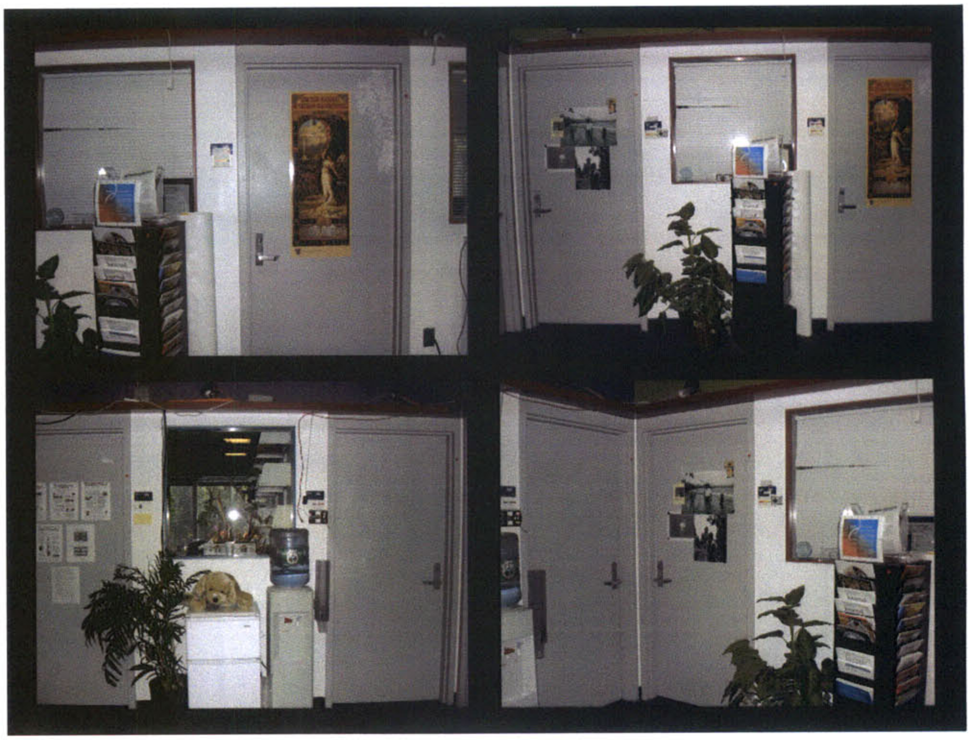
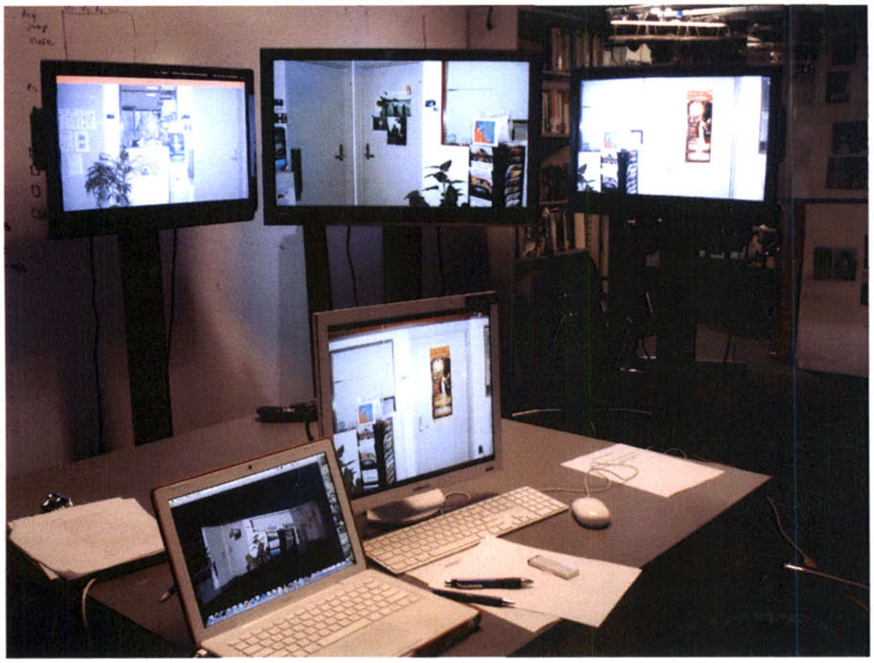
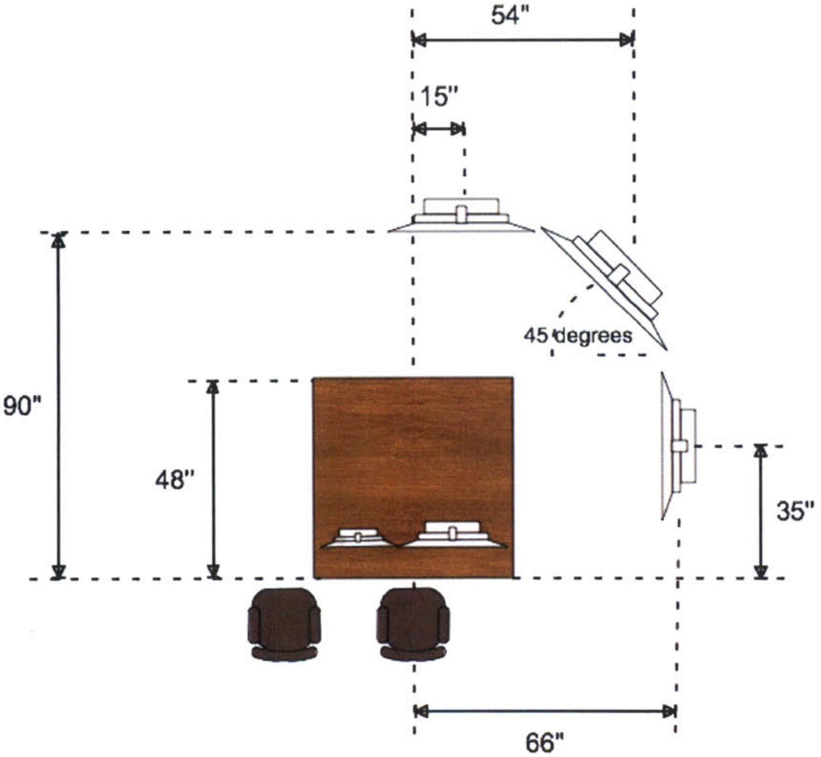


Fig. 5-10 Tiled view, top-down, left to right order: images 3, 4, 1 and 2

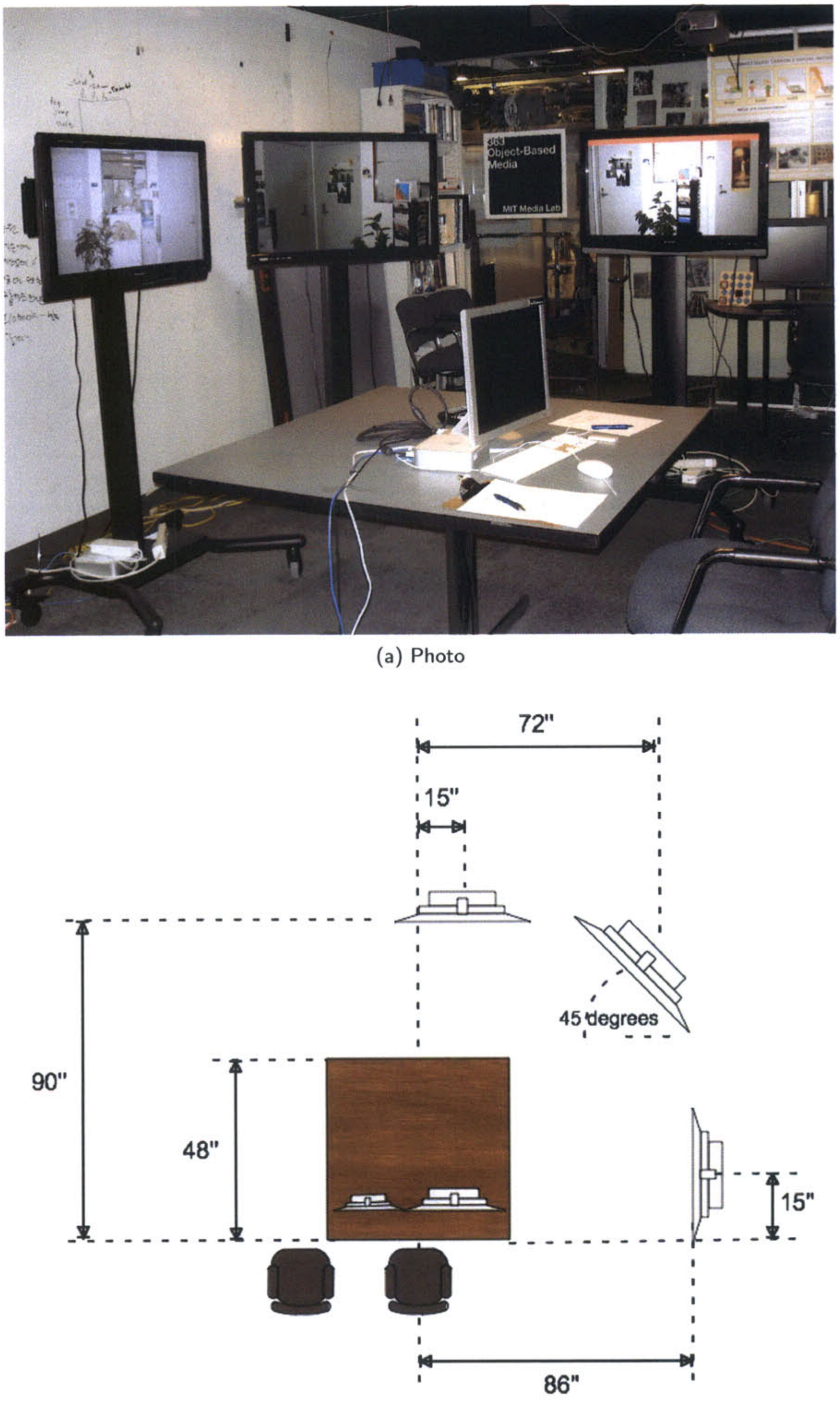


(a) Photo



(b) Top-view diagram

Fig. 5-11 Experiment setup repositioned by subjects - monitors positioned close to each other



(a) Photo

(b) Top-view diagram

Fig. 5-12 Experiment setup repositioned by subjects - monitors positioned far apart

6

Conclusion

In this chapter, we present an overall conclusion of the project, which mostly covers discussions about the system implementation and the major lessons learned from the experiments. Additionally, we present possible future directions of research.

6.1 Major Remarks

As previously mentioned, uCom's vision is to enhance the feeling of being present in two locations simultaneously. This vision is implemented by enabling real-time awareness of what happens in a remote site. The system explores the idea of "spatial displays", i.e., it displays images according to their spatial relationship. It actually uses video to create a visual representation that is coherent with the layout of a remote space.

The main contributions of uCom derive from (1) the intermediate 3D representation of a remote space, (2) a simple user interface that allows users to portray remote views on fairly any displays available locally, and (3) the ability to incorporate fairly any video acquisition and display equipment. First, It is noteworthy the novelty of rendering a graphical visualization resembling a 3D collage of live video feeds. The contribution also derives from using this type of 3D interface for a remote awareness application. The combination of user interface's navigation controls to switch camera viewpoints and the portrayed transition path between them have clearly suggested to have helped users on sensing how the images relate in the 3D space. Second, uCom also creates a multi-display environment, in which users can easily control the image shown in each displays, and how to physically position the displays in order to fit their needs. The user interface is an easy-to-use remote control interface that allows users to choose which available view to be shown in each display. Third, the system design can make use of fairly any video acquisition and display equipment, whether professional or off-the-shelf and low-cost devices.

uCom's evaluation assesses questions related to the geometric cues of a remote space. Ultimately, we investigate to what extent uCom improves users' awareness of a remote location's layout using the remote views. We evaluated both the 3D image-based model with respect to a tiled arrangement of the images, and user's interactions with the multiple displays available in a uCom room. Our studies recruited subjects who were not familiar with the scene, i.e., they had never been to the portrayed location so they had no clue of its spatial layout. Comparing user's reactions to the two display formats representing a remote space, most subjects seemed to prefer the tiled arrangement rather than the 3D image-based model. Yet, most of them could effectively create a mental model of the remote space using the 3D image-based model. Most subjects claimed that the distortions in the 3D image-based model made it harder for them to identify objects in the scene to be used as references for a mental model of the space's layout. Only a few subjects claimed that the 3D model easily enabled them to create a mental model of the remote space, particularly due to the navigation controls that followed the arrangement of the images. Concerning mapping views of the remote space to displays in the local environment, our experiments suggested a potential use for constant remote monitoring but not necessarily immersion in the remote space. In fact, most subjects suggested system use cases related to remote surveillance and monitoring. Its is noteworthy that most users tried to enable the images in the monitors so as to match the geometric position of the portrayed scene. Additionally, the experiments indicate that the brightness from the monitors raised concerns about its potential for distracting users from other activities. Another relevant observation is that most subjects took advantage of the possibility of moving the displays. They made positive remarks about the flexibility the system provides, allowing displays to be moved to fit users' needs and changes to the furniture layout.

6.2 Future Work

Future work on uCom will possibly involve the following main issues: extending the user studies to evaluate more general use case scenarios, and adding extra features to the system. First, the user studies will possibly focus on evaluating: (1) the impact on the assignment of views to each displays when multiple users present in each uCom space, (2) the use of uCom in different environments rather than workspaces, and (3) the impact of peripheral displays on users' attention, and (4) privacy related issues. Second, we foresee implementing additional system features, such as (1) enabling audio communication through uCom, (2) supporting special image acquisition hardware, (3) improving navigation features in the 3D image-based model.

Additionally, we intend to allow users to freely engage with the system in order to create appropriate forms of use and applications. Those currently unforeseen uses might drive more meaningful iterations of the system design.

6.2.1 Evaluation of other use case scenarios

We would like to evaluate how multiple users in one uCom space negotiate how to display the views of a remote space. As all subjects present in one space share the same peripheral displays, there is a potential complexity of having several observers in one room that need to agree on which images should be portrayed in each of the displays. We expect to identify conflicts between the needs of the group versus the needs of individuals, and propose guidelines or additional system features that can mitigate them. Additionally, as in real environments it might not be possible to position displays and orient control spaces to satisfy each user's preference, we should evaluate the penalty on performance if either or both user's preferences are not met; or how willing are users to engage with displays if they are not in individually owned, among related research questions.

It is noteworthy that uCom's current implementation and user tests are focused in a work related environment. Yet we believe the uCom concept could be applied to connecting any two architectural spaces that have similar purposes. For instance, to connect the homes of family members who live apart, or even to serve a mutually agreed purpose, such as connecting a patients' hospital rooms to their homes.

Finally, we intend to evaluate the peripheral displays' impact on users' attention. For that, we foresee creating situation-specific tasks that users should primarily perform, while the displays in the uCom space portray situations that might interfere on users' attention. Perhaps we could explore different kinds of scene portrayals, wether enabling face-to-face or in-context view from users in a remote space. In addition, we intend to create specific use case scenarios to draw conclusions on privacy and symmetry concerns.

6.2.2 Additional system features

We foresee the addition of audio capabilities to uCom in order to enable users to directly communicate through our system. Yet audio would require rethinking the current system design. Questions remain open on how to create a correspondence between multiple video views and the audio connections. It would require positioning microphones and speakers accordingly. Nevertheless, we believe the addition of audio is a natural next step to convey immersion between remote spaces.

We also consider the possibility of using special image acquisition hardware, such as cameras with 3D range sensing, UV or thermal-IR and other capabilities. They could be used to portray remote users' motion activity in the remote space. They could symbolically represent the remote activity level, rather than using solely video and audio. For that, we could create alerts to attract remote users' attention. Additionally, we could use eye-tracking to monitor the attention cost of observing a peripheral display with images of a remote location.

Our experiments suggest the need for two major improvements to the 3D image-based model of a remote scene: (1) to compensate for image distortions, and (2) to enhance the navigation controls. First, most of the image distortions visible in our 3D image-based model are introduced by having images rendered on non-rectangular shaped quadrangles. Some of these distortions could be minimized with projective texture mapping techniques, which allow textured images to be projected onto a scene as if by a slide projector. Second, we plan to improve the navigation controls that enable users to switch their point of view among the available cameras' estimated viewpoints. They currently allow users only to switch between cameras according to their relative position in one direction, the horizontal axis. We plan to improve those controls by enabling users to switch between cameras along at least two sets of axes - vertical and horizontal - in all possible directions, i.e., left to right, right to left, up to down, and down to up.

Bibliography

[1] Bundler. http://phototour.cs.washington.edu/bundler/.

[2] Cave automatic virtual environment (cave). http://www.evl.uic.edu/pape/CAVE/.

[3] Cisco telepresence. http://www.cisco.com/en/US/netsol/ns669/networking_solutions_solution_segment_home.html.

[4] Google street view. http://maps.google.com/help/maps/streetview/.

[5] Hp halo. http://h71028.www7.hp.com/enterprise/us/en/halo.

[6] icom. http://web.media.mit.edu/~stefan/hc/projects/icom/.

[7] Microsoft photosynth. http://livelabs.com/photosynth/.

[8] S. Agamanolis. New technologies for human connectedness. *SPECIAL ISSUE: Ambient intelligence: the next generation of user centeredness*, 12(4):33–37, July + August 2005.

[9] S. Agamanolis and V.M. Bove, Jr. Multilevel scripting for responsive multimedia. *Multimedia*, 4(4):40–50, October - December 1997.

[10] R.M. Baecker. *Readings in Groupware and Computer Supported Cooperative Work: Software to Facilitate Human-Human Collaboration.* Morgan Kaufman Publishers, 1993.

[11] H.H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M.E. Goss, J. MacCormick, K. Yuasa, W.B. Culbertson, and T. Malzbender. Computation and performance issues in coliseum, an immersive videoconferencing system. In *MM*, pages 470–479, 2003.

[12] J. Baldwin, A. Basu, and H. Zhang. Panoramic video with predictive windows for telepresence applications. In *IEEE International Conference on Robotics and Automation*, pages 1922–1927, 1999.

[13] A. Borning and M. Travers. Two approaches to casual interaction over computer and video networks. In *CHI*, pages 13–19, 1991.

[14] M. Brown and D. Lowe. Unsupervised 3d object recognition and reconstruction in unordered datasets. In *International Conference on 3D Digital Imaging and Modeling*, pages 56–63, 2005.

[15] C. Cool, R.S. Fish, R.E. Kraut, and C.M. Lowery. Iterative design of video communication systems. In *CSCW*, pages 25–32, 1992.

[16] P. Dourish. Re-space-ing place: "place" and "space" ten years on. In *CSCW*, pages 299–308, 2006.

[17] P. Dourish and V. Bellotti. Awareness and coordination in shared workspaces. In *CSCW*, pages 107–114, 1992.

[18] P. Dourish and S. Bly. Portholes: Supporting awareness in a distributed work group. In *CHI*, pages 541–547, 1992.

[19] M. Fiala. Structure from motion using sift features and the ph transform with panoramic imagery. In *Canadian Converence on Computer and Robot Vision (CRV)*, 2005.

[20] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model tting with applications to image analysis and automated cartography. In *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pages 726–740. Morgan Kaufmann, 1987.

[21] R.S. Fish, R.E. Kraut, R.W. Root, and R.E. Rice. Evaluating video as a technology for informal communication. In *CHI*, pages 37–48, 1992.

[22] W. Gaver, T. Moran, A. MacLean, L. Lovstrand, P. Dourish, K. Carter, and W. Buxton. Realizing a video environment: Europarc's rave system. In *CHI*, pages 27–35, 1992.

[23] W. Gaver, A. Sellen, C. Heath, and P. Luff. One is not enough: Multiple views in a media space. In *INTERCHI*, pages 335–341, 1993.

[24] S.J. Gibbs, C. Arapis, and C. Breitenender. Teleport - towards immersive copresence. *Multimedia Systems*, (7):214–221, 1999.

[25] R.I. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, 2004.

[26] J. Hollan and S. Stornetta. Beyond being there. In *CHI*, pages 119–125, 1992.

[27] S. Ikeda, T. Sato, and N. Yokoya. High-resolution panoramic movie generation from video streams acquired by an omnidirectional multi-camera system. In *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 155–160, 2003.

[28] B. Johanson, A. Fox, and T. Winograd. The interactive workspaces project: Experiences with ubiquitous computing rooms. *Pervasive Computing Magazine*, 1(2), April-June 2002.

[29] P. Kauff and O. Schreer. An immersive 3d video-conferencing system using shared virtual team user environments. In *CVE*, pages 105–112, 2002.

[30] M. Lourakis and A. Argyros. The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. `http://www.ics.forth.gr/~lourakis/sba`, 2004.

[31] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, September 2004.

[32] A. Majumder, W.B. Seales, M. Gopi, and H. Fuchs. Immersive teleconferencing: A new algorithm to generate seamless panoramic video imagery. In *ACM Multimedia*, 1999.

[33] M. Mantei, R. Baecker, Sellen, W. A., Buxton, T. Milligan, and B. Wellman. Experiences in the use of a media space. In *CHI*, pages 203–208, 1991.

[34] J. Mulligan, X. Zabulis, N. Kelshikar, and K. Daniilidis. Stereo-based environment scanning for immersive telepresence. *Transactions on Circuits and Systems for Video Technology*, 14(3), March 2004.

[35] M. Nacenta, S. Sallam, B. Champoux, S. Subramanian, and C. Gutwin. Perspective cursor: Perspective-based interaction for multi-display environments. In *CHI*, pages 289–298, 2006.

[36] N Negroponte. *Being Digital.* Vintage, 1996.

[37] D. Nguyen and J. Canny. Multiview: Spatially faithful group video conferencing. In *CHI*, pages 799–808, 2005.

[38] Y. Nomura, L. Zhang, and S.K. Nayar. Scene collages and flexible camera arrays. In *Eurographics Symposium on Rendering*, 2007.

[39] K. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita. Multiparty videoconferencing at virtual social distance: Majic design. In *CSCW*, pages 385–393, 1994.

[40] C. Plaue and J. Stasko. Present & placement: Exploring the benefits of multiple shared displays on an intellective sensemaking task. In *GROUP*, pages 179–188, 2009.

[41] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *SIGGRAPH*, pages 179–188, 1998.

[42] N. Snavely, Seitz S. M., and R. Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH*, pages 835–846, 2006.

[43] N. Snavely, Seitz S. M., and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, December 2007.

[44] Noah Snavely. *Scene Reconstruction and Visualization from Internet Photo Collections.* PhD thesis, University of Washington, December 2008.

[45] N. A. Streitz, J. Geissler, T. Holmer, S. Konomi, C. Muller-Tomfelde, W. Reischl, P. Rexroth, P. Seitz, and R. Steinmetz. i-land: An interactive landscape for creativity and innovation. In *CHI*, 1999.

[46] R. Stults. Media space. Xerox PARC Technical Report, 1986.

[47] S. Sunil Arya, M.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu. An optimal algorithm for approximate nearest neighbor searching xed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[48] R. Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Computer Vision*, 2(1), December 2006.

[49] B. Takacs, A. Beregszaszi, and G. Komaromi-Meszaros. Panocast: A panoramic multicasting system for mobile entertainment. In *IEEE International Conference on Information Visualization (IV)*, 2007.

[50] J.C. Tang, E.A. Isaacs, and M. Rua. Supporting distributed groups with a montage of lightweight interactions. In *CSCW*, pages 23–34, 1994.

[51] L. Teodosio and M. Mills. Panoramic overviews for navigating real-world scenes. In *Multimedia*, 1993.

[52] K. Tomite, K. Yamazawa, and N. Yokoya. Arbitrary viewpoint rendering from multiple omnidirectional images for interactive walkthroughs. In *16th International Conference on Pattern Recognition*, pages 987–990, 2002.

[53] L. Vincent. Taking online maps down to street level. *Computer*, 40(12), December 2007.

[54] W. C. Wen, H. Towles, L. Nyland, G. Welch, and H. Fuchs. Toward a compelling sensation of telepresence: Demonstrating a portal to a distant (static) office. In *IEEE Visualization*, pages 327 – 333, 2000.

[55] D. Wigdor, C. Shen, C. Forlines, and Balakrishnan R. Effects of display position and control space orientation on user preference and performance. In *CHI*, pages 309–318, 2006.