Supporting collaborative practices across wall-sized displays with video-mediated communication

Ignacio Avellino

université PARIS-SACLAY

NNT : 2017SACLS514

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Supporting collaborative practices across wall-sized displays with video-mediated communication

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°580 Sciences et technologies de l'information et de la communication

Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Orsay, le 12 décembre 2017, par

Ignacio Avellino

Composition du Jury :

Jean-Claude Martin Professeur, Université Paris-Sud (– LIMSI)	Président
Carman Neustaedter Professeur, Simon Fraser University (– SIAT)	Rapporteur
Myriam Lewkowicz Professeur, Université de Technologie de Troyes (–*Tech-CICO*)	Rapporteur
Albrecht Schmidt Professeur, Ludwig-Maximilian University of Munich	Examinateur
Michel Beaudouin-Lafon Professeur, Université Paris-Sud (– HCC, ExSitu)	Directeur de thèse
Cédric Fleury Maître de Conférences, Université Paris-Sud (– HCC, ExSitu)	Co-Directeur de thèse

Thèse de doctorat

ABSTRACT

Collaboration can take many forms, we might sit side by side to work on an artifact, stand around a table to manipulate shared objects, or stand in front of a large display to visualize big datasets. Technology has long provided support for these practices through many devices: desktop computers let two people work side by side on digital objects, tabletops let groups of people gather around shared data, and wall-sized displays support visualizing and manipulating large digital data sets. Their traits determine how people use them in co-located collaboration. But when collaborators are located remotely, to what extent does technology support these activities?

In this dissertation, I argue that the success of a telecommunications system does not depend on its capacity to imitate co-located conditions, but in its ability to support the collaborative practices that emerge from the specific characteristics of the technology. I explore this question using wall-sized displays as a collaborative technology. Wall-sized displays are large and can present massive data sets at a high resolution, two traits that establish the behaviors that take place in these spaces. Notably, people physically navigate data, moving close to and far away from the display instead of zooming in and out, and walking towards objects instead of panning. These opportunities shape how people collaborate: they can simultaneously and independently navigate data, indicating far objects to each other through gestures; as well as perform tightly-coupled work, physically navigating data together while talking about it.

To explore technological support for remote collaboration across wall-sized displays, I started by observing collaborators perform their daily work at a distance using low-fidelity and technological prototypes. These observations showed how video of the remote collaborator is used during collaboration, mainly for deictic gestures and for discussions, and guides the rest of the work in the dissertation. I then found through experiments that people can accurately interpret remote deictic instructions and direct gaze when performed by a remote collaborator through video, even when this video is not placed directly in front of the observer. This suggests that we can move video during remote collaboration while people physically navigate the space, without disrupting communication. Based on these findings, I built *CamRay*, a telecommunication tool that realizes collaboration across large interactive spaces. *CamRay* uses an array of cameras to capture users' faces as they physically navigate data on a wall-sized display, and presents this video in a remote display on top of existing content. This tool is designed to explore collaboration needs across wall-sized displays and how to support them. I use *CamRay* to observe how pairs perform two different collaborative tasks at a distance. Based on these observations, I propose two ways of displaying video: *Follow-Local* and *Follow-Remote*. In *Follow-Local*, the

video feed of the remote collaborator follows the local user, and in *Follow-Remote* it follows the remote user. I perform two experiments to evaluate how these video behaviors support different aspects of collaboration. I find that *Follow-Remote* preserves the spatial relations between the remote speaker and the content, supporting pointing gestures. *Follow-Local* enables virtual face-to-face conversations, supporting representational gestures. Finally, I summarize these findings to inform the design of future systems for remote collaboration across wall-sized displays, and reflect on the considerations that designers of telecommunication systems should take when adding support for remote communication across collaborative technologies. I present three implications for the design of telecommunication systems. First, video feeds of remote collaborators do not necessarily need to be placed in front of the observer for remote gestures to be accurately understood. Second, video placed in congruence with content better supports remote pointing gestures. And third, video placed in the focus area better supports representational gestures.

RÉSUMÉ

La collaboration entre plusieurs personnes peut prendre plusieurs formes, telles que se placer côte à côte pour travailler ensemble sur un objet, autour d'une table pour manipuler des objets ou debout devant un grand écran pour visualiser un grand nombre de données. La technologie soutient depuis longtemps ces différentes pratiques au travers de nombreux dispositifs: les ordinateurs de bureau permettent à des individus de travailler côte à côte sur des objets numériques, les tables interactives permettent de se rassembler autour de données partagées, et les murs d'écrans permettent de visualiser et de manipuler facilement de grands ensembles de données numériques. Les caractéristiques propres de ses dispositifs déterminent la manière dont les gens les utilisent pour collaborer dans le même espace. Mais lorsque la collaboration doit se faire à distance, est-elle aussi bien assistée par la technologie ?

Dans ce travail, je soutiens l'idée selon laquelle le succès d'un système de télécommunications ne dépend pas de sa capacité à imiter une collaboration colocalisée, mais dans sa capacité à faciliter les pratiques collaboratives découlant des caractéristiques spécifiques de la technologie. J'explore cet argument en utilisant un mur d'écrans en tant que technologie collaborative. Les murs d'écrans offrent un espace d'affichage important et permettent de visualiser de grandes quantités de données avec une haute résolution, deux caractéristiques qui conditionnent le comportement des collaborateurs. Ceux-ci peuvent notamment naviguer physiquement face aux données, se rapprocher ou reculer par rapport à l'écran au lieu de zoomer ou dézoomer, et se déplacer d'objets en objets au lieu de faire défiler le contenu. Ces traits déterminent la manière dont les individus collaborent: ils peuvent naviguer simultanément et indépendamment, tout en indiquant des objets aux autres par des gestes; effectuer un travail conjoint en naviguant physiquement dans les données tout en les discutant.

Dans l'optique d'explorer le support technologique pour la collaboration distante entre murs d'écran, j'ai commencé par observer des collaborateurs effectuer leur travail quotidien à distance en utilisant des prototypes de basse fidélité et technologiques. Ces observations ont montré comment la vidéo du collaborateur distant est utilisée pendant la collaboration, principalement pour les gestes déictiques et pour les discussions, et guide le reste du travail de cette dissertation. Ensuite j'ai montré à l'aide d'expérimentations contrôlées que les utilisateurs peuvent interpréter avec précision les instructions déictiques à distance et le regard direct quand un collaborateur à distance est affiché par une vidéo, même si celle-ci n'est pas placée directement devant l'observateur. Ceci suggère que nous pouvons déplacer un flux vidéo dans la collaboration à distance pendant que les collaborateurs naviguent physiquement l'espace, sans perturber la com-

munication. À partir de ces résultats, j'ai créé *CamRay*, un outil de télécommunication qui permets la collaboration dans de grands espaces interactifs. *CamRay* utilise une rangée de caméras pour enregistrer le visage des utilisateurs lorsqu'ils parcourent physiquement les données le long de l'écran et présente cette vidéo sur un autre mur d'écrans distant par dessus le contenu existant. Cet outil permet d'étudier les besoins nécessaires à la collaboration à distance entre murs d'écrans et d'explorer différentes solutions pour comment la faciliter. Dans cette thèse, j'utilise *CamRay* pour observer comment des paires d'utilisateurs effectuent deux tâches collaboratives différentes à distance. En observant la façon dont cet outil répond aux besoins de la collaboration, je propose deux possibilités pour afficher la vidéo: *Follow-Local* et *Follow-Remote*. Avec *Follow-Local*, le flux vidéo de l'utilisateur distant suit l'utilisateur local, tandis qu'avec *Follow-Remote*, le flux vidéo suit l'utilisateur distant. Je présente les résultats de deux expériences évaluant comment ces deux stratégies prennent en charge différentes séquences de la collaboration. Plus particulièrement, je montre que *Follow-Remote* préserve les relations spatiales entre le collaborateur à distance et le contenu de l'écran, créant ainsi la possibilité de désigner les objets par des gestes de pointage, tandis que *Follow-Local* facilite les conversations grâce à un face-à-face virtuel qui transmet plus facilement la communication gestuelle. Finalement, je me base sur ces résultats pour guider la conception de futurs systèmes de communications à distance entre murs d'écrans, et dégager des considérations à suivre lorsque des capacités de communication à distance sont ajoutées à de nouvelles technologies. Je présente trois implications pour la conception de systèmes de télécommunication. Premièrement, les flux vidéo de collaborateurs distants n'ont pas nécessairement besoin d'être placés devant l'observateur pour que les gestes à distance soient bien compris. Deuxièmement, placer la vidéo en congruence avec le contenu facilite les gestes de pointage à distance. Et troisièmement, la vidéo placée dans la zone de focus facilite les gestes de représentation.

ACKNOWLEDGMENTS

First and foremost, I thank my two advisors, Michel Beaudouin-Lafon and Cédric Fleury. Michel, thank you for guiding me and mentoring me ~~in~~**for** the past three years~~;~~**,** for teaching me to think critically, work independently, and for setting a good example ~~in~~**on** countless occasions. Oh, and for the endless English corrections. Cédric, thank you for teaching me how to look for new perspectives during frustrating road blocks, for sharing your research experiences, and for being there until the last minute in every deadline.

I also want to thank Wendy E. Mackay for being an unofficial advisor. Wendy, thank you for taking a great deal of your time and sharing your knowledge on conducting observations, experiments, doing data analysis, and for teaching me the subtleties of the research world. I deeply thank the three of you for the kindness with which you always answered my questions, teaching me something new every day.

I thank my jury members Carman Neustaedter, Myriam Lewkowicz, Albrecht Schmidt and Jean-Claude Martin for taking time out of their busy schedules to read my thesis and provide extremely valuable comments.

INRIA and Paris-Sud funded my doctoral studies, I thank these institutions for their trust in me.

Everyone in the ExSitu team provided a unique, joyful and productive working environment during my studies. Thank you Sarah Fdili Alaoui for sharing not only your knowledge, but your amazing spirit. You are a role model to me. Thank you Fanis Tsandilas for all your advise on how to do proper statistics. Thank you Nolwenn Maudet and Jessalyn Alvina, my thesis writing companions, for always keeping the morale up! Thank you Nolwenn for your friendship, being so cheerful and supportive at home and at work. Marianela Ciolfi, Carla Griggio and Germán Leiva, thank you for providing a shoulder in desperate times during conference deadlines. I especially thank Germán for dealing with the complexities of setting a live broadcast of my thesis defense! My parents are forever grateful. Oleksandr Zinenko and Ghita Jalal, thank you for being the senior mentors that introduced me to the team. Philip Tchernavskij and Midas Nouwens, thank you for all the intellectually stimulating discussions.

I want to thank everyone from the i10 lab at RWTH Aachen University, who mentored me during my master studies. I especially thank Chatchavan Wacharamanotham who has always been an endless source of knowledge and inspiration .. and a great friend.

Thank you Rabah Taouachi for you love and daily support during the production of this work. Sorry for being so difficult at times.

Lastly, I thank my parents for giving me and my brother the most essential tool for leading an independent life: an education. I know it took a heavy heart, so thank you for giving me the freedom to pursue my studies abroad.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

"Hitherto, the fact that face-to-face contact has almost always been the most satisfactory form of communication has been a fundamental constrain to society"

— John Short [94]

Collaboration is at the center of our activities. Technology provides support for various collaborative practices, such as siting side-by-side in front of a desktop, gathering around data using a tabletop display, or presenting information to a group using a digital whiteboard. When these practices take place remotely, we turn to telecommunication technologies. These technologies are oftentimes used simply as a means to get audio-video signals across two distant locations, overlooking collaborative aspects. I am interested in studying the integration of remote communication technology and collaborative technology to support existing co-located collaborative practices in a remote setting. I created a system that adds remote collaboration capabilities to two distant wall-sized displays and allows investigating how to support a variety of tasks at a distance. By isolating parts of the communication process, I performed studies and observed how people change the way they communicate when the telecommunication technology changes. Oftentimes we tend to see technology as having a passive mediation *role; I promote the view that technology plays a larger, more active role in communication, a* modulation *role.*

Different people master different skills. To solve complex problems that involve a mixed set of disciplines, people come together and collaborate. Collaboration can take many forms: people may sit side by side to work on a shared artifact, gather around tables to hold discussions and manipulate objects, or stand in front of a group to share and discuss ideas. Existing systems provide digital support to these diverse collaborative practices, such as connected mobile devices, desktop computers, interactive tabletops or large screens, each of which has distinct characteristics that make it fit for specific tasks. Thus, the choice of a particular technology reflects the needs of a group during collaboration. Two people may choose a mobile device to quickly share media [77], or a desktop setting to edit a digital document, as this supports side-by-side conversations [108]. Groups may choose a tabletop display for discussing and manipulating data, as its size, height and horizontal layout support gathering around it [60]. Also, groups may choose a wall-sized display for visualizing highly-dense and large datasets, as its size and resolution support visualizing massive amounts of information simultaneously and at different scales [8].

The characteristics of each technology provide unique opportunities for co-located collaboration.

But in today's globalized world, it is nothing but commonplace to find remote group members, connected through telecommunication systems. So far, we have largely solved remote *communication* from a technical perspective, providing high-quality audio-video links and the ability to share data in real time. But in doing so, we have often neglected remote *collaboration*.

I argue that the success of a telecommunication system does not depend on its capacity to imitate co-located conditions, but on its ability to support the collaborative practices that emerge from the specific characteristics of the technology. Simply sending audio and video signals in an attempt to simulate being in the remote space is not sufficient. Short [94] identified the limitations of imitating co-located situations as early as in 1976: *"However, even a full-colour life-size three-dimensional motion-picture transmission of the complete body of the distant person by some futuristic television system, although a complete replication of the face-to-face situation in its reproduction of non-verbal and verbal cues, would not, according to the Social Presence approach, be exactly the same as face-to-face contact as long as the interactors are aware that they were separated. The knowledge of physical separation may suffice to make the telecommunicated interaction more like other telecommunicated interactions, rather than like face-to-face"* (pp. 158-159). It is thus pointless to try and imitate the real world, as from the moment collaborators are aware that technology mediates communication, their perception is different than in a real-world situation. Systems should instead go *Beyond Being There* [57] in order to both overcome the limitations of the technology and enable configurations that are not possible when people are co-located. In this dissertation, I explore this argument using wall-sized displays as a collaborative technology.

Wall-sized displays are large vertical surfaces that can display vast data sets at ultra-high resolutions [10]. They radically change the way people interact with data, as physical navigation lets them walk towards objects as opposed to panning them on a screen, and walk towards and away from objects as opposed to zooming [8]. This makes wall-sized displays especially fit for a variety of tasks such as data visualization, visual information searching, sense making and data manipulation. In low-level data visualization and navigation tasks, wall-sized displays significantly outperform desktop displays that use pan and zoom for navigation [7]. In visual information search, having a larger display lets people scan and find objects more easily and can increase efficiency [74]. In sense-making tasks, participants use the large space to distribute cognition and tend to arrange information spatially as opposed to annotating it and switching among maximized windows [2]. In data manipulation tasks, as difficulty increases when classifying more complex data, wall-sized displays outperform desktop displays [75].

The characteristics of wall sized-displays do not only benefit single users, but also bring new opportunities for co-located collaborative practices. Collaborators take advantage of the space and navi-

gate through data by physically moving their bodies [64], which they find fast, intuitive and natural [63]. They can independently and simultaneously interact with information [8], improving tasks such as problem solving [62], but at the same time they can engage in tightly-coupled collaboration, increasing efficiency during data manipulation tasks [76]. Collaborative use of wall-sized displays has been studied from a co-located perspective, but as people are more and more often located remotely, it is important to study how to implement remote collaboration across these spaces. I thus focus on the following research questions:

- How can we keep the benefits of co-located collaboration using wall-sized displays, but in a remote setting?
- How can we overcome the limitations that arise from mediating communication with technology?
- How can we enable new configurations that are not possible in co-located settings?

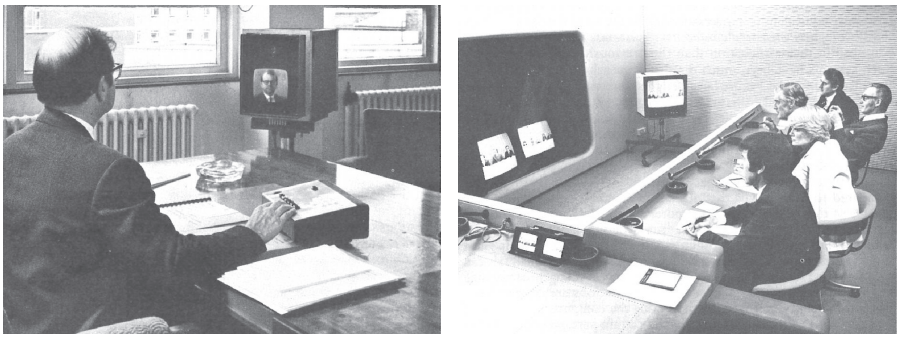
(a) A loudspeaking telephone (British Post Office LST4).

(b) The remote Meeting Table (Dover House, London).

Figure 1: Two remote audio link systems from the 1960's: (a) a loudspeaking telephone (British Post Office LST4) and (b) the remote Meeting Table (Dover House, London). Images from Short, 1976 [94].

Telecommunication technologies have provided audio and video links to implement remote collaboration for a long time. Telephone links support remote group conversations as far back as the 1960's (Figure 1 a) and even round-table style conferences where a speaker and a microphone replace a remote participant, giving a spatial reference in the local space (Figure 1 b). Videophones support one-to-one conversations over the desk also as far back as the 1970's (Figure 2 a), even with the ability to capture—besides faces—physical documents on a desk by pushing a switch to flip a mirror. Remote group meetings also have had support in large spaces through video links as far back as the 1970's (Figure 2 b).

Research has studied how telecommunication technologies, like the ones just mentioned, can be used to support collaborative practices at a distance while taking into account the activities that the technology is used for. For instance, Tang and Minneman's VideoDraw [106] and



(a) An experimental Viewphone (British Post Office).

(b) The Confravision studio (Euston Tower, London).

Figure 2: Two remote video link systems from the 1960's: (a) an experimental Viewphone (British Post Office) and (b) the Confravision Studio (Euston Tower, London). Images from Short, 1976 [94].

VideoWhiteboard [107] explore how to achieve remote shared drawing, taking into account that people need to understand each other's hand gestures. Sellen and colleagues' Hydras [92] focus on remote conferencing, and explore where to place remote collaborators' audio and video feeds to account for their need to interpret each other's gaze.

In the next chapter, I present a more detailed overview of systems that support collaborative practices at a distance and are designed to take advantage of the underlying characteristics of the technology they connect. Although some systems that provide communication across wall-sized displays exist, previous studies have not addressed the exploration of how to leverage the characteristics of this technology to support existing co-located collaborative practices, but at a distance. This exploration is necessary. Sellen [91], as early as the 1990's, argued that *"research has not really addressed the issue of how the specifics of the design of different videoconferencing systems might affect behavior"*, and Bordia [86] recognized that *"it is alarming how little we know about the effects of computer-mediated communication on interpersonal influence, persuasion, impression formation and management, power relations, and personal perception"*.

To study collaboration across large interactive spaces, I designed and built *CamRay*, a telecommunication system that connects remote people using wall-sized displays. *CamRay* uses an array of cameras to capture users' faces as they physically navigate data in a wall-sized display. It presents this video on a remote display over existing content. This system enabled me to explore the needs for communication across wall-sized displays, and to support existing practices during co-located collaboration, in a remote setting. *CamRay*'s architecture is modular and flexible, thus making it extendable. It can support researchers in the future to unlock the full potential of these spaces during remote collaboration and to understand how people use them.

1.1 THESIS STATEMENT

It is commonplace to find systems that simply establish audio-video links between remote locations to realize communication at a distance. I argue that telecommunication systems need to support the collaborative practices that emerge from the specific characteristics of technology if they want to accomplish remote *collaboration* as opposed to just *communication*. In this dissertation, I explore this argument using wall-sized displays as a collaborative technology. These displays are characterized by their large size and high-resolution, which promote physical navigation of data. I implement a system that enables the study of remote collaboration across wall-sized displays, and use it to show that user movement is not a weakness during collaboration, but an action to leverage when displaying video of remote collaborators that can benefit collaboration. I show that we can support various collaborative practices that emerge from the use of wall-sized displays at a distance. In this exploration, I also discover that by establishing what's possible during communication, this system does not only *mediates* communication but also *modulates* it, by changing the way people communicate.

1.2 RESEARCH APPROACH

Although a promising technology, wall-sized displays are not yet commonly found in industry. The difficulty of conducting in-situ observations of real-world experts who master this tool in their daily work, makes it hard to discover the needs for collaboration from a user's perspective. I thus start my exploration by conducting qualitative studies to observe how people would use the technology when collaborating remotely. I perform a laboratory observation where I simulate a remote collaboration scenario using a low-fidelity paper prototype. This observation informs the design of a first technological prototype that I use to study in a second observation of how people use video of remote participants during communication. I then use these results to build *CamRay*, a telecommunications systems that captures people's faces as they move in front of a wall-sized display, and presents their video on a remote wall-sized display.

With this system, I perform quantitative studies to evaluate the technology. I conduct two experiments using *CamRay* to explore the trade-offs of displaying video following each collaborator's movement, and how to support gestural and conversational communication.

1.3 THESIS OVERVIEW

The rest of this thesis is organized as follows:

Chapter 2 summarizes personal communication and previous work on computer-mediated communication. I discuss how previous sys-

tems have implemented remote collaboration for different technologies, and position my work in the context of wall-sized displays.

Chapter 3 presents initial observations. I build a low-fidelity paper prototype that simulates two remote locations by dividing one large display into two, and observe two collaborators assemble a conference presentation together. I then prototype a telecommunications system that can capture video of each participant and display it in the other simulated location. I observe two co-authors work on a shared literature classification task, and conclude with design guidelines for the next steps of this system.

Chapter 4 presents an initial exploration into communication across wall-sized displays, where I study direct eye gaze perception and the accuracy of remote deictic gesture interpretation.

Chapter 5 describes *CamRay*, a system that uses an array of cameras to capture users' faces as they physically navigate through data, and can display their video in a remote tiled display. This system is a tool for exploring collaborative practices in these spaces and how to support them.

Chapter 6 presents observations using *CamRay* during two different collaborative tasks. Informed by the findings of the previous observations, I propose two video behaviors that leverage each collaborators' position to display video: *Follow-Local* and *Follow-Remote*. In *Follow-Local*, the video feed of the remote collaborator follows the local user, and in *Follow-Remote* it follows the remote user. I conclude by presenting my findings on how each technique supports different moments in collaboration.

Chapter 7 reports on an experiment that explores how to support deictic gestures in remote collaboration using *CamRay*. I use a data manipulation task that relies on the use of pointing gestures to evaluate the trade-offs of each video behavior. The results show that using the remote collaborator's position to display video (*Follow-Remote*) preserves the spatial relations between the remote speaker and the content, supporting the use of pointing gestures.

Chapter 8 reports on an experiment where I explore how to support face-to-face communication in remote collaboration using *CamRay*. This time, I create a task that operationalizes hand gestures that occur during conversation, to evaluate the trade-offs of each video behavior. The results show that using the local collaborator's position to display video (*Follow-Local*) enables virtual face-to-face conversations, supporting the use of representational gestures.

Chapter 9 summarizes the results of the two studies and situates them within the theory of personal communication. I present guidelines on the design of remote communication systems for large interactive spaces. Finally, I discuss some unexplored aspects of remote communication in this thesis, leaving space to further research in this area.

2

COMMUNICATION, TECHNOLOGY MEDIATION AND LARGE INTERACTIVE SPACES

> *"As multimedia becomes an integral part of collaborative systems, we must understand how to design such systems to support users' rich set of existing interaction skills, rather than requiring people to adapt to arbitrary constraints of technology-driven designs."*
>
> — Isaacs & Tang [59]

This chapter presents previous work on telecommunication systems and how they support collaborative practices. I introduce the mechanisms people use to communicate, focusing on gestures and distinguishing between deictic and representational gestures. I then discuss technology-mediated communication and how they support existing practices. I introduce wall-sized displays and discuss their benefits for single users and for collaboration. Finally, I discuss how, so far, previous work has only focused on enabling communication as in face-to-face situations when connecting two large interactive spaces. In my work, I study how to overcome the limitations of communication technology, giving support to existing collaborative practices but at a distance, as well as enabling configurations that are not possible when co-located.

In this literature review, I firstly introduce the various mechanisms that people use to communicate, focusing on gestures. I then make the distinction between deictic and representational gestures, which I use throughout this dissertation. I discuss technology-mediated communication, presenting existing work that supports daily co-located collaborative practices without technology, as well as with technology. I introduce wall-sized displays and present its benefits, focusing on their ability to enable physical navigation of data. I discuss how these spaces shape collaboration and present systems that have enabled remote communication across them. Finally, I position my work within the reviewed literature.

2.1 COMMUNICATION

Colin Cherry [25] defines communication as *"the physical signals whereby one individual can influence behavior of another"*. Though short, this definition shows that the essential part of communication is to influence the behavior of another person.

Communication can occur in many forms, using spoken words or sounds, the body or even physical artifacts. Gutwin [50], for example,

distinguishes several mechanisms that people use to gather information from others, three of which involve the speaker (direct communication, indirect productions and consequential communication) and two the surrounding environment (Feedthrough and Environmental feedback). In direct communication, people intentionally communicate information, which can be verbal but also through gestures [105] or deictic references [109]. With indirect productions, people communicate through actions that are not explicitly directed at other members of the group but are intentionally public [35, 53]. In consequential communication, people attend to the other persons' bodies in the workspace, as their activities reveal information [90]. Feedthrough is when information can be gathered by observing the effects of other people's actions on artifacts [33]. Environmental feedback is when people perceive indirect effects of another person's actions in the workspace, such as observing that a value has changed which indicates that someone initiated an action.

These three mechanisms involve people using their bodies to communicate information. Kendon [66] also recognizes that people continuously inform each other, willingly or not, about their intentions, interests, feelings and ideas by means of visible bodily action.

In summary, the body, and in particular the production of gestures, plays a large role in communication.

2.1.1 *Gestures and Speech*

Gestures, according to Kendon [66], are *"actions that have the features of manifest deliberate expressiveness"*. The use of gestures is common practice: Tang et al. [105] found that people performed gestures about 35% of the time they produced hand gestures during group design sessions.

Social science researchers, notably Argyle [4], McNeal [80] and Kendon [66], have long studied the combination of speech and gestures in relation to a task and shown how complexly intertwined they are. When gestures can both facilitate speech production [70] and interpretation [87].

In Human–Computer Interaction, Bekker et al. [12] observed how speech and gestures are intertwined when analyzing face-to-face conversation in design projects, Tang et al. [105] when observing collaborative drawing by small groups, and Kirk et al. [68] when studying different remote gesture technologies in a physical task.

Gestures can add to speech, for instance to help explain something, or they can even replace speech completely. Speech in itself is highly dependent on the mutual knowledge collaborators share prior to their interaction, their *common ground* [28, 30]. The process of accumulating common ground, by exchanging evidence of what collaborators understand or not during conversation is called *grounding* [30]. Common ground determines the choices of language during conversation. Isaacs & Clark [58] observed for example that experts rely on spe-

cialized vocabulary during interactions with novices, which increases conversation efficiency.

When two people are co-present (i.e. in the same room), they rely on the fact that they know what the other person can see and hear to formulate utterances. This leads to the use of *demonstrative references*, which, according to Clark et al. [29] *"require an accompanying gesture for its complete interpretation"*.

2.1.1.1 *Two Types of Gestures: Deictic and Representational*

There are many classifications of gestures in the literature. Kendon [66] identifies at least four major ones: Efron, 1941 [37]; Ekman & Friesen, 1969 [38]; Kendon, 1972 [65] and McNeal, 1992 [80]. Fussel et al. [42] also discuss these classifications, and point out that they all distinguish between *pointing* and *representational* gestures. *Pointing* gestures are used to refer to task objects and locations. *Representational* gestures are used to represent the form of task objects and the nature of actions to be used with those objects. Previous work in HCI used this distinction, e.g. Bekker et al.'s study of design teams [12] or Tang et al.'s work on mixed presence groupware [102].

Fussel et al. [42] further distinguish three types of representational gestures: (1) *iconic representations* to show what an object looks like, (2) *spatial* gestures to show distances and (3) *kinetic* gestures to demonstrate how an action should be performed.

In this dissertation, I only distinguish between pointing and representational gestures as they are the most relevant to the situations I studied.

2.2 TECHNOLOGY-MEDIATED COMMUNICATION

Technology has long provided support for communicating at a distance. In the early days telephones allowed voice to be transmitted across long distances. Later video links became available in experimental phones, then computers, and today they are a standard feature of smartphones. In 1976, Short [94] noted the benefits of video, but that its transmission in the future would incur a high cost, because the necessary bandwidth to transmit video is ten times that of audio. However, video has become ubiquitous in communication today, at no extra cost.

This is largely because, as it is the case in face-to-face conversation, the ability of collaborators to see each other during technology-mediated communication has many advantages and plays an important role in communication. Isaacs & Tang [59] studied the benefits of video over audio during remote communication, and found that video allows to express understanding and agreement, forecast responses, enhance verbal descriptions, give purely nonverbal information, express attitudes through posture and facial expression, and manage extended pauses. Veinott et al. [111] showed that seeing each other's faces in collaborative tasks improves the negotiation of common ground, as opposed to using non-video media. Monk & Gale [81]

observed that having mutual gaze awareness provides an alternative to non-linguistic channels for awareness of a remote person's understanding.

When people are located remotely, collaborators not only pay attention to their partners but also to their workspace, as when co-located. Clark & Krych [27] identify five perceptual regions that people monitor during video-mediated collaboration: three relate to the person (face, body and voice) and the other two to objects (workspaces and shared scenes).

2.2.1 *Supporting Collaborative Practices in Daily Activities Remotely*

A number of previous systems have implemented remote collaboration by taking into account the activities that they support. In this section, I start by giving examples of systems that support remote collaboration for activities that do not need any technology when they are co-located, and discuss how they take into account user practices in their design. I then present different technologies, the collaborative practices that emerge from their characteristics, and how systems support remote collaboration beyond simple audio-video links.

2.2.1.1 *Shared Drawing*

Shared drawing was one of the first activities to receive attention.

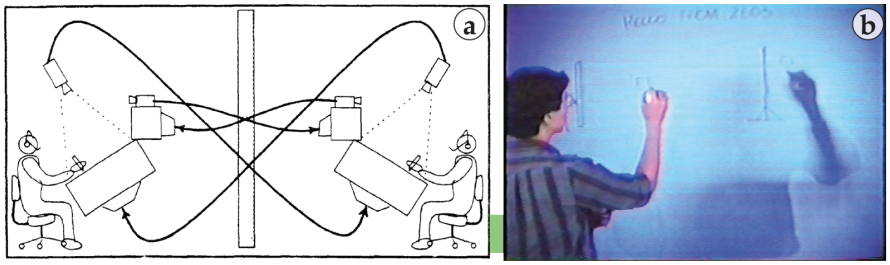


Figure 3: Tang & Minneman's (a) *VideoDraw* [106] and (b) *VideoWhiteboard* [107]. An image of the remote user is overlaid on top of content to form a *reference space* [21]. Image from [21].

In *VideoDraw* [106] (Figure 3 a) and *VideoWhiteboard* [107] (Figure 3 b), Tang and Minneman overlaid the video feed of remote participants on top of shared drawings to support hand gestures and to give a sense of spatial relationship among the collaborators and the drawing space. Bly and Minneman [15] continued this effort with *Commune*, adding shared cursors, the ability to mark drawings and to write text, with the goal of supporting natural uses of the drawing surface.

These systems enable shared drawing at a distance, supporting practices such as gazing and pointing at parts of drawings by overlaying the remote collaborator's face and arms on top of content. Performing hand gestures to augment speech enabled the use of deictic expressions in these systems. They implement what Buxton refers to as the a *reference* space [21], the combination between shared *per-*

son and *task* spaces [22]. The *task* space is where each collaborator has shared objects that they can perceive and manipulate. The *person* space is where collaborators can see each other and perceive their facial expressions, voice, gaze and body language among other facial and bodily cues.

2.2.1.2 *Conferencing*

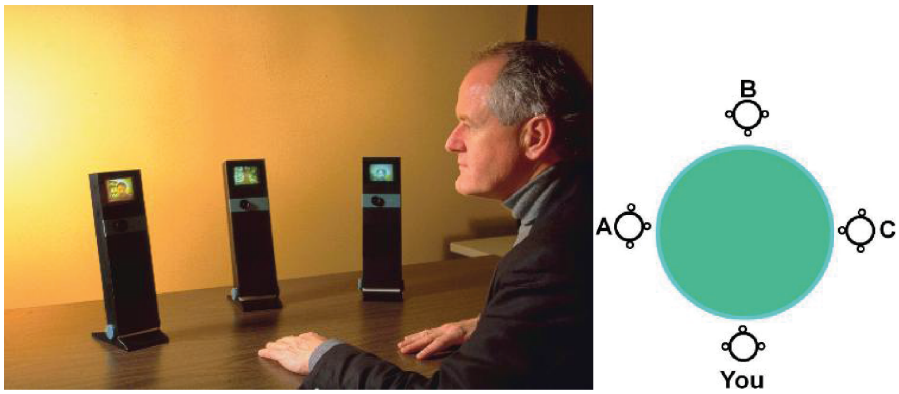


Figure 4: Sellen et al.'s [92] *Hydras* conferencing system, a remote conferencing system that preserves gaze direction to remote partners by placing their video feeds in a way that reflects their original position. Image from [21].

Sellen et al. built the *Hydras* [92], where video and audio feeds of remote participants in a multi-party conversation are placed such that they reflect the original participants' position. The Hydras convey gaze direction and lets understand who is being targeted in a conversation (Figure 4).

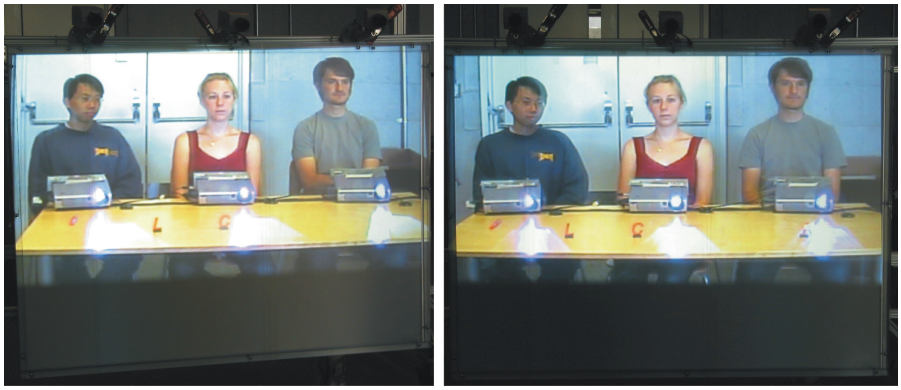


Figure 5: Nguen et al.'s *Multiview* conferencing system. Each participant sees the remote participants' gaze direction as in a face-to-face situation. Image from [82].

Nguyen & Canny [82] later built *Multiview*. Each site has one camera and one projector per participant. They capture and display participants from each remote person's perspective. As a result, each person sees the remote participants' gaze direction with spatial faithfulness, as if there was a mirror between the two sites. This system

preserves non-verbal cues during remote conferencing, notably gaze (Figure 5).

Non-verbal cues have important functions during conversation, such as providing awareness of mutual attention, controlling the floor, giving feedback, making illustrations, emblems or expressing interpersonal attitudes [3]. By supporting correct gaze interpretation in remote conversations, The *Hydras* and *Multiview* support practices such as yielding the floor to someone by looking at them, or providing clear indications of who is being addressed.

2.2.2 *Supporting Collaborative Practices with Technology Remotely*

2.2.2.1 *Tabletops*

Tabletops have characteristics that lead to two main behaviors during collaboration: specific coupling styles and territoriality.

Coupling styles are related to the physical arrangements of the members in a group.

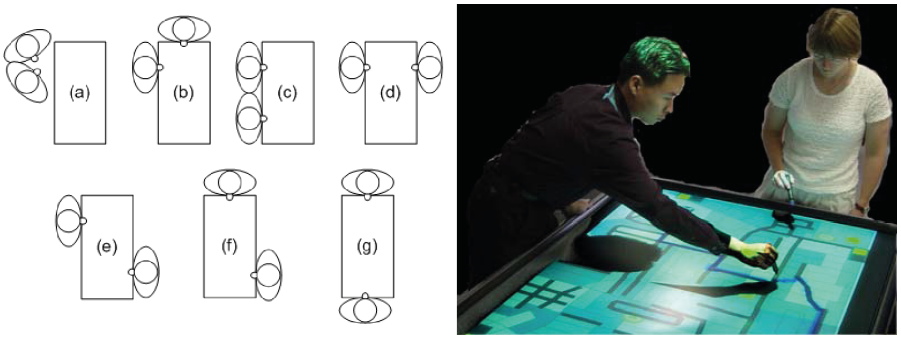


Figure 6: Tang et al.'s study on collaborative coupling over tabletop displays. (left) seven position arrangements around the table coded during the study: (a) together, (b) kitty corner, (c) side by side, (d) straight across, (e) angle across, (f) end side, and (g) opposite ends. (right) experiment setup. Images from [103].

Tang et al. [103] performed two observational studies where they identified six styles of collaborative coupling: both collaborators working on the same problem, in the same area, one working with the other viewing while engaged in the task, both working on the same problem in different areas, one working and the other viewing but not engaged, one working and the other fully disengaged, and both working on different problems. Participants also adopted several position arrangements, frequently and fluidly engaging and disengaging with group activity (Figure 6). They conclude that *"groups use tighter coupling styles when working together closely, preferring common, global views"*. Isenberg et al. [60] performed observations in a collaborative sense-making task and extended Tang et al.'s work, adding two more collaboration styles. Their work shows how collaboration styles vary according to how loosely or tightly coupled pairs work. They found that 73% of their participants spent 70% of their time on average in

close collaboration, while the other 27% spent 60% of their time working in parallel.

Territoriality refers to how people partition the shared space around a tabletop.

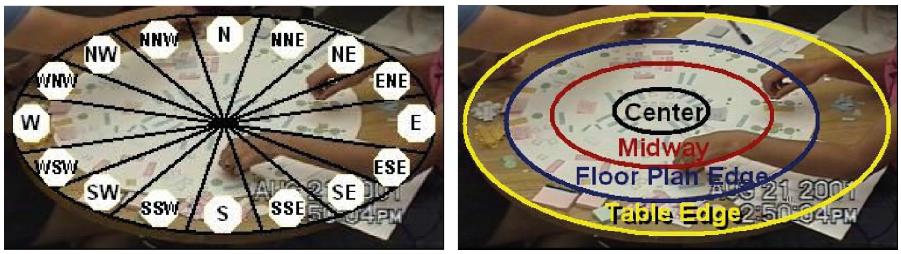


Figure 7: Scott et al.'s territoriality study in tabletops, directional and radial zones. Images from [89].

Tang [105] observed during pen and paper activities that people reserve the area immediately in front of them as a personal space. Kruger et al. [72] observed the same phenomena in interactive tabletop displays, and also that object orientation relates to how people establish personal and group spaces and how they signal ownership of objects. Scott et al. [89] conducted two observational studies, a preliminary one to observe tabletop interactions in a casual environment, and a second one where pairs worked on layout planning activities (Figure 7). They found that in shared tabletops, people use three types of territories to help coordinate interactions: personal, where they reserve an area and resources for their personal use; group where they perform activities with others; and storage to keep unused objects.

Given these collaborative practices that emerge in tabletops, how much support have systems provided for them in remote collaboration?

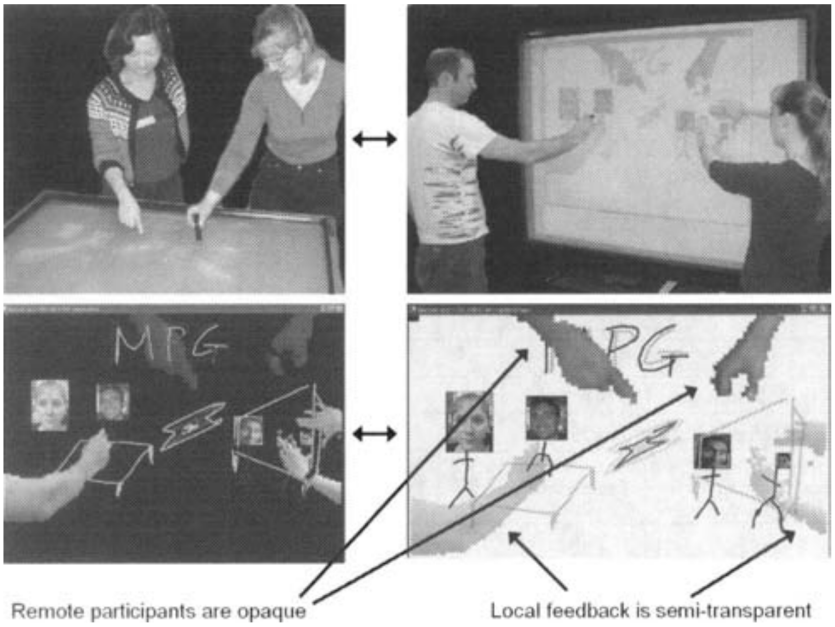


Figure 8: Tang et al.'s *VideoArms* system. A Mixed Presence Groupware session using this system. Participants arms are captured and overlaid on top of content remotely. Image from [102].

Tang et al. [102] created *VideoArms* to eliminate presence disparity in remote settings (Figure 8). Presence disparity arises when tabletop users have different perceptions of their local and remote counterparts. *VideoArms* mitigates this disparity by capturing video of arms on top of tabletops and displaying them remotely according to their original spatial location. The authors observe that people perform and understand rich task-related gestures with this system. Also, they spend considerable amounts of time watching each other to understand the state of the task. This suggests that *VideoArms* supports awareness of other participants location, and that coupling styles and territoriality could perhaps be supported, although authors do not investigate this.

In a similar spirit, Tang et al. [101] observed that overlaying shadows of the remote collaborators' arms over shared content increased awareness of their actions, as they are more visible than telepointers. Gutwin et al. [49] showed that workspace awareness, *"the up-to-the-moment understanding of another person's interaction with the shared workspace"* is necessary to collaborative coupling. Thus we can assume that since remote projected arms increase awareness of collaborator's actions, they support collaborative coupling.

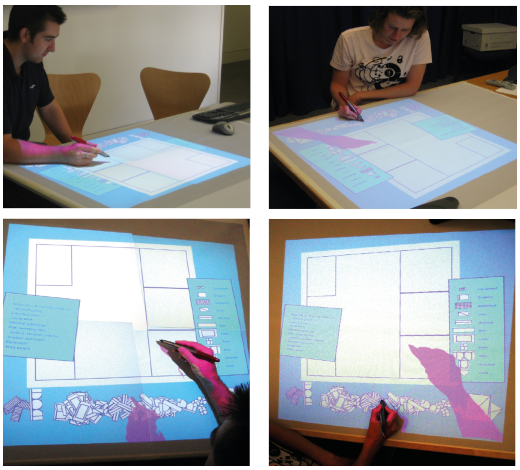


Figure 9: Tuddenham et al.'s studies of territorial coordination and workspace awareness in remote tabletop collaboration. Image from [110].

Tuddenham et al. [110] provide more solid evidence for this claim. The authors used projections of remote arms in a co-located and remote collaborative task using tabletops. They found no differences in collaborative coupling between the remote and local conditions, which suggests that projecting remote arms supports this practice in similar ways. For territoriality though, they do observe differences. In the local condition participants partitioned the space according to who was nearest. But in the remote condition this didn't happen, participants were not hesitant to work across each other, and to take objects from in front of each other. They find that remote collaborators did not coordinate territorially as co-located collaborators did.

It seems then that projecting remote collaborators arms does support some collaborative practices (the use of gestures, coupling styles), but not others (territoriality). On the one hand, it seems that *VideoArms* goes *being being there* [57] by avoiding the problems that come with territoriality. Nonetheless, it might be useful to support territorialy in remote systems.

2.2.2.2 *Large Interactive Spaces*

In this dissertation, I focus on large interactive spaces that contain large displays, the next section presents the benefits of using these spaces, how co-located collaboration is shaped by these benefits, and how previous systems have realized remote collaboration.

2.3 WALL-SIZED DISPLAYS

Wall-sized displays [10] can present massive data sets at a high resolution (Figure 10). This creates new opportunities for interaction with data. In this section, I review previous work that shows the benefits of these spaces, focusing on physical navigation.

Figure 10: A wall sized display with 6 people collaborating in pairs on the CHI2013 conference program. Image copyright: Inria.

2.3.1 *Benefits*

A large visualization surface brings benefits to a number of tasks, ranging from everyday computer use to spatial, sense-making and spatial orientation tasks.

Figure 11: Czerwinski et al.'s *DSharp* display. Image from [32].

Switching windows on wall-sized displays can be performed more efficiently than in regular screens. Czerwinski et al. [32] conducted a study comparing their *DSharp* display (46.5″) to a desktop display (15″), and found that the first increased productivity and satisfaction for complex tasks that rely on multiple windows (Figure 11). Bi and Balakrishnan [14] compared a 16′ × 6 (6144 × 2034 pixels) wall display using projections with single and dual desktop monitors in a week-long study. They concluded that the enhanced peripheral awareness that large screens provide, facilitates tasks with multiple windows. Grudin [47] also found focal and peripheral regions of multiple (desktop) displays to help users manage multiple windows.

Daily work with computers is also improved when performed in large displays. Ball & North [6] observed five users during six months

use a large tiled display (a 3 × 3, 17″ LCD titles) for their everyday work. In their study, people took advantage of the screen space, spatially positioning applications to assign them meaning, improving task switching and increasing their awareness for secondary tasks. In Czerwinski et al.'s [32] study of everyday work, participants increased their productivity using a large screen and were highly satisfied.

Spatial tasks were studied by Tan and colleagues, showing that they notably improve when using large displays. Tan et al. [97] showed in an first experiment that wall-sized displays outperforms a normal size screen for spatial orientation, when holding the visual angle constant by adjusting the viewing distance to each of the displays. Tan et al. [98] later extended this results, and showed in an experiment that users are more effective at performing 3D virtual navigation tasks on large displays. The authors further argue in a later study that more immersive environments encourage egocentric rotations, leading to improved performance [99]. Finally, Tan et al. [96] used the *Dsharp* display in an experiment, and showed that large displays increases 3D navigation performance in females, due to the presence of optical flow cues

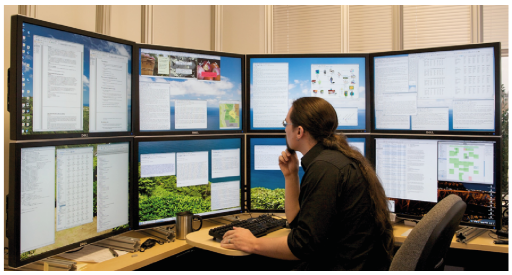


Figure 12: Andrew et al.'s study of sense-making tasks in large displays. People assign objects meaning using spatial locations. Image from [2].

Sense-making tasks also benefit from increased screen real estate. Andrews et al. [2] showed that a large screen can become part of users' distributed cognitive processes, providing a form of external memory and a semantic layer (Figure 12). This allows people to assign objects meaning using spatial locations.

Search and comparison tasks in Information-Rich Virtual Environments (IRVE) also benefit from large displays. Ni et al. [83] compared three types of displays to evaluate the effect of resolution: a desktop, a back projection using one projector and a high-resolution projection based on an array of projectors. Their results show that large high-resolution displays improve navigation as well as search and comparison.

2.3.1.1 *Physical Navigation*

Physical navigation is a common practice that comes as a consequence of the large space available in wall-sized displays. This practice has many advantages, both for single-user work and for collaboration.

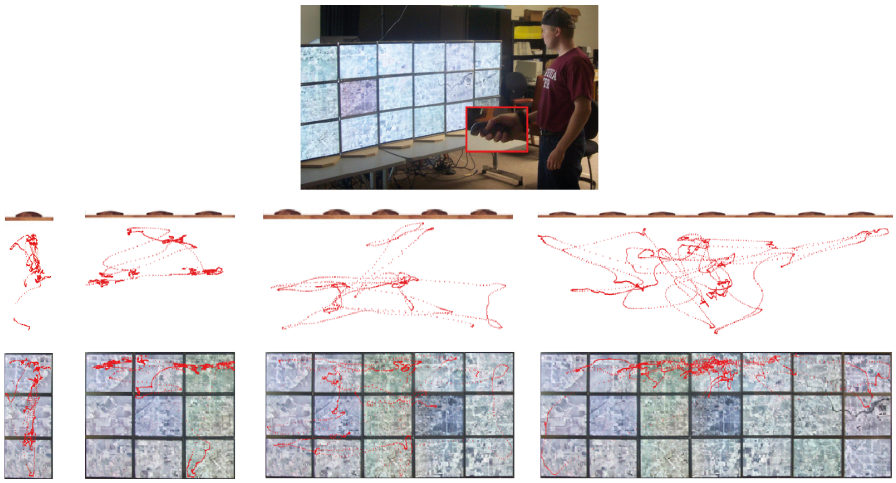


Figure 13: Ball et al.'s study of physical navigation in wall-sized displays. As the screen space available increases, physical navigation and performance increase, and virtual navigation decreases. Image from [8].

Ball et al. [7] in a first study showed that higher resolution displays that use physical navigation significantly outperform smaller displays that use pan and zoom navigation. The authors later controlled for screen size to determine its relation to the amount of physical and virtual navigation as well as task performance in an experiment [8]. They showed that as screen space increases, virtual navigation decreases and user performance improves (Figure 13).

Yost et al. [117] explored the perceptual scalability of information visualizations for large displays. They conducted an experiment using an 8 × 3 tiled display of LCD monitors with a resolution of 1280 × 1024 each. They vary display size, using sizes with a sufficient number of pixels to be within (2560 × 768 pixels), equal to (5120 × 1536 pixels), or beyond visual acuity (10240 × 3072 pixels). The results show that even with the navigation costs incurred from physically moving, there is an increase in performance, and sometimes in accuracy, because of the additional data that could be displayed in larger spaces.

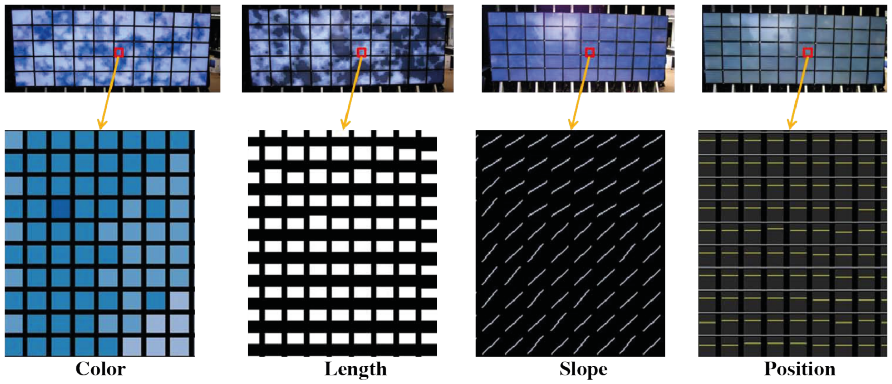


Figure 14: Endert et al.'s study on visual encodings in wall-sized displays (color, length, slope and position). Image from [40].

Endert et al. [40] showed that the visual encodings such as color, length, slope or position support physical navigation (Figure 14). They compared large (10 × 5 20″ LCD monitors, 16000 × 6000 pixels) and small (one 20″ LCD monitor of 1600 × 1200 pixels) displays and showed that color, among other encodings, promotes physical navigation more effectively when exploring a data space. They acknowledge that the choice of a visual encoding should consider the way in which it aggregates data. For instance, color is highly noticeable, length aggregates as luminance and slope as texture when seen from afar.

Bezerianos & Isenberg [13] studied the perception of visual variables on wall-sized displays. They conduct two studies using a 5.5m × 1.8m tiled display (32 30″ LCD monitors, 20480 × 6400 pixels in total) and measure the accuracy of their judgments when standing far from or close to the display, and when moving freely. They find that moving was as accurate as static comparisons from afar, and recommend encouraging viewers to stand further back from the display when conducting quantitative comparison tasks. This study shows that the ability for people to move in front of large displays does not come at the cost of miss interpreting visual variables.

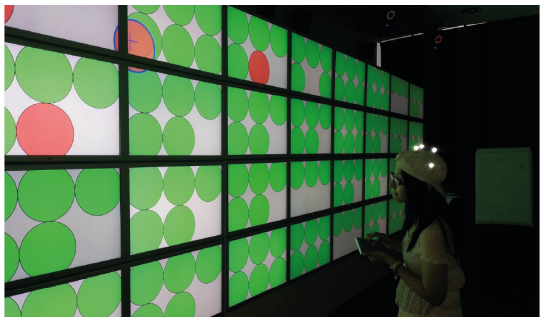


Figure 15: Liu et al.'s study on visual physical navigation vs. virtual navigation. A participant using a trackpad to classify targets. Image from [75].

Finally, Liu et al. [75] compared physical navigation in front of a wall-sized display (8 × 4 30″ LCD monitors, 20480 × 6400 pixels in total) to virtual navigation using pan-and-zoom on a desktop (one 30″ LCD monitor of 2560 × 1600 each) (Figure 15). Their experiment consisted on classifying information through physical manipulation, with various levels of difficulty of the task. The authors find that as difficulty increases, larger screens provide higher efficiency.

In summary, previous work shows that physical navigation increases as screens become larger and is preferred by users, which translates into benefits in a wide range of tasks. As we will see now, these benefits also extend to collaborative work.

2.3.2 *Wall-Sized Displays and Collaboration*

Wall-sized displays are especially fit for collaboration, as they can easily accommodate multiple users.

Large displays have been used to support co-located collaboration in different settings, for instance for face-to-face meetings [39], office work [95], military command [36], and high school education [19].

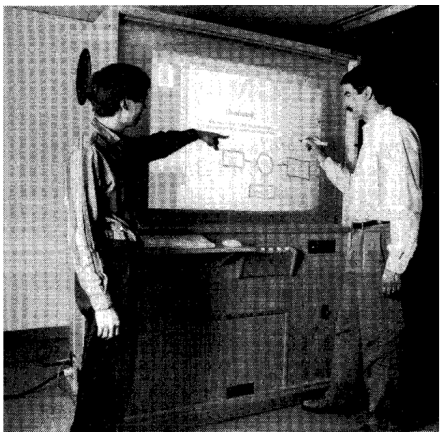
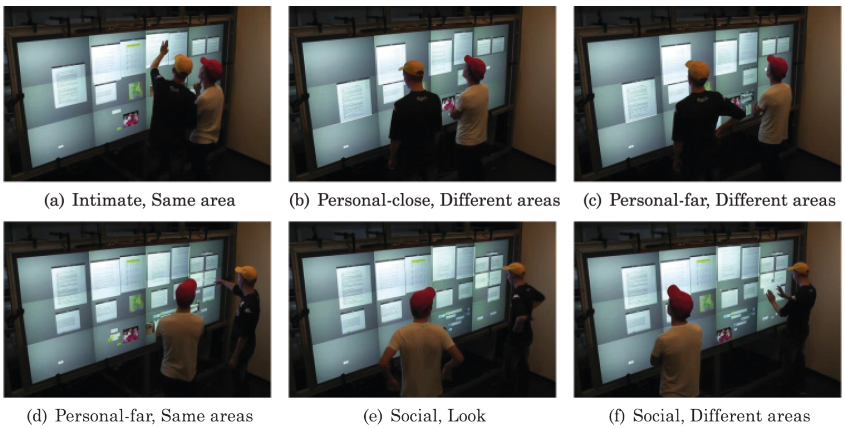


Figure 16: Elrod et al.'s *Liveboard*, one of the first large interactive vertical surfaces. Image from [39].

Elrod et al.'s *Liveboard* [39] (Figure 16) was probably one of the first systems that let more than one person interact with a vertical surface of one million pixel using a pen. It was designed to support a variety of group tasks such as meetings using a whiteboard, slide show presentations, games and even programming. This systems was designed to support group work and while the authors conducted informal surveys on its use, they did not specifically study how its use affects collaboration.

Guimbretière et al.'s Interactive Mural [48] is an interactive display with 9 megapixels, designed for brainstorming sessions. Their goal is to provide digital support for situations where people work collaboratively with a large collection of information on a physical wall. Although they support collaboration, the authors did not study the effects of their system on collaborative practices.



(a) Intimate, Same area (b) Personal-close, Different areas (c) Personal-far, Different areas

(d) Personal-far, Same areas (e) Social, Look (f) Social, Different areas

Figure 17: Jakobsen & Hornbæk's study on a collaborative problem-solving task. The figure shows the 6 coupling styles found: (a) Intimate, Same area; (b) Personal-close, Different area; (c) Personal-far, Different areas; (d) Personal-far, Same areas; (e) Social, Look; (f) Social, Different areas. Image from [62].

Jakobsen & Hornbæk [62] studied a collaborative problem-solving task with a 24 megapixel, 2.8m × 1.2m multitouch display. Participants had to work with several documents to find a hidden plot. Authors identified six collaboration styles (Figure 17), they found that proximity plays a role in how tightly coupled participants work and that simultaneous input reduces the need for coordination in loosely coupled work. They also found that people did not fight for space (territoriality) but shared it evenly, switching between joint and parallel work. Finally, they observed, just as with tabletops, that collaborators' physical position reflects how closely they work.

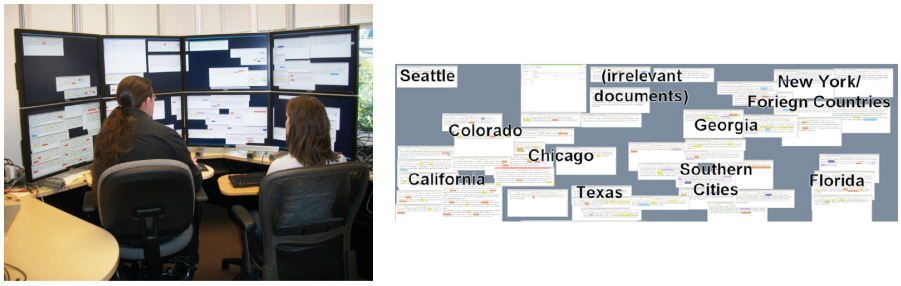


Figure 18: Bradel et al.'s study of collaborative sense-making tasks on large displays. (left) The experimental setup. (right) An example of geographical document clustering. Image from [16].

Bradel et al. [16] studied collaboration in a sense-making task. Participants sat in front of a 4 × 2 grid of 30″ LCD 2560 × 1600 pixel monitors (10.240.3200 pixels in total) curved display and interacted with documents using a mouse, with the goal of predicting a future terrorist attack (Figure 18). Participants arranged information spatially, establishing common ground between pairs about the meaning of the spatial arrangements they created. This arrangement led to closer collaboration and better performance, as people could find documents more easily.

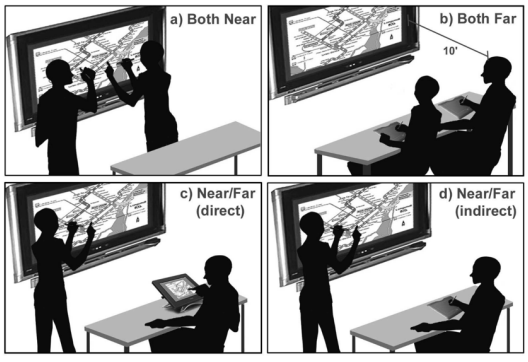


Figure 19: Hawkey et al.'s study of proximity on collaboration using a large display. The image shows the four experimental conditions. Image from [52].

Hawkey et al. [52] examined the impact of proximity on the effectiveness and enjoyment of co-located collaboration using a large display. Participants worked in pairs in a route-planning task based on subway routes in four conditions: both near, both far, one near and

one far with direct or indirect input (Figure 19). Proximal collaboration was preferred in general; and, when working at a distance, direct input is preferred as the information is visible in the input space but awareness of each other's actions is hindered as gestures are less visible.

Figure 20: Liu et al.'s study of collaborative data manipulation task using a large display. Three typical participant behaviors: (a) working separately, (b) showing a container and (c) telling the partner what to do. Image from [76].

Liu et al. [76] studied co-located collaboration in a data manipulation task on a 5.5m wall-sized display (8 × 4 30″ LCD monitors, 20480 × 6400 pixels) (Figure 20). Pairs worked in five different collaboration styles, four of which result by combining loose and close collaboration with provided or not provided shared interaction, and a baseline condition where there is no communication. The results show the trade-offs of communication: although more communication may hinder efficiency, it is still beneficial to users as they perceive it as more efficient, more enjoyable and are more engaging. Also, the authors found that shared interaction techniques support collaboration, as users are more efficient, travel shorter distances, and prefer this condition.

In summary, as with tabletops (Section 2.2.2.1), wall-sized displays have unique traits that bring new opportunities for collaboration. In the next section I present systems that have looked at remote work across large interactive spaces.

2.3.3 *Wall-Sized Displays and Remote Collaboration*

Previous work has developed systems that provide support for remote communication across spaces that contain large screens.

Figure 21: (a) Beck et al.'s immersive group-to-group telepresence and (b) Maimone et al.'s encumbrance-free telepresence system. Images from [11] and [79].

One approach is to create 3D virtual representations of the remote collaborators. Beck et al. [11] use depth cameras to capture users, and present them in a remote location inside a shared 3D virtual scene (Figure 21 a). This provides an immersive experience, where users are "transported" to remote locations. Multiple users are supported by using perspectively correct stereoscopic images. Maimone et al. [79] also use multiple depth cameras to capture users, this time with a focus on creating a clear 3D reconstruction of the remote person, filling its holes and smoothing it (Figure 21 b). These approaches focus on the complex technology that is needed to recreate co-located situations with realism, but they do not explore collaboration needs and how these tools satisfy them.

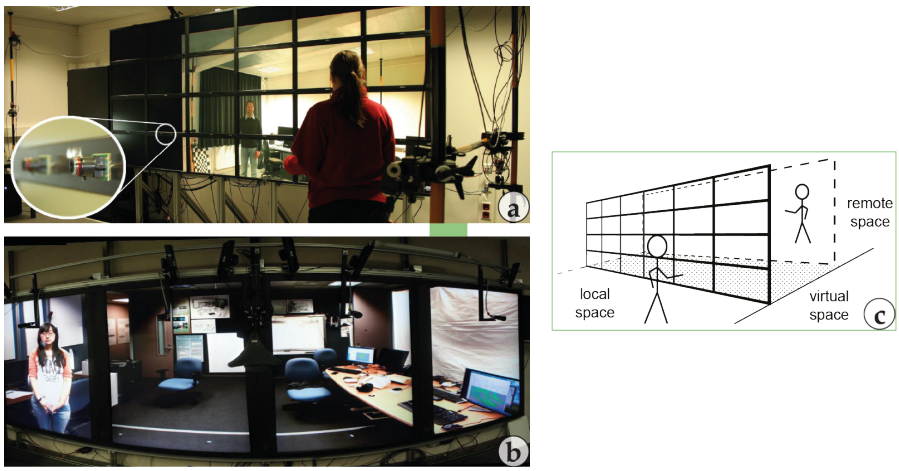


Figure 22: (a) Willert et al.'s 2D array of cameras for wall-sized displays and (b) Dour et al.'s room-sized informal telepresence system. Two systems that implement (c) the transparent window metaphor. Images from [115] and [34].

Another approach for remote collaboration is through the glass-window metaphor (Figure 22 c). Willert et al. [115] connect two remote wall-sized displays by capturing video of a remote site through a 2D array of cameras, mounted on the screens' bezels, and display this video remotely with a parallax effect as the observer moves (Figure 22 a). Dou et al. [34] realize this metaphor by using RGB cameras to capture a background scene and depth cameras to capture people in the foreground. They segment the person in the foreground using the 3D image data, and overlay it onto a panorama of the background. This approach can achieve the sense of direct eye contact as people are captured using the camera in front of them. Using multiple cameras to create the extended window metaphor has the advantage of supporting user movement; however, using the full display for video does not leave room for shared digital content, which is necessary for collaboration. This approach effectively corrects for eye gaze (Figure 22 b). Although, these systems support communication across two distant sites, they do not consider shared digital information, as they display remote video using the entire available screen space. Also, they do not focus on studying communication.

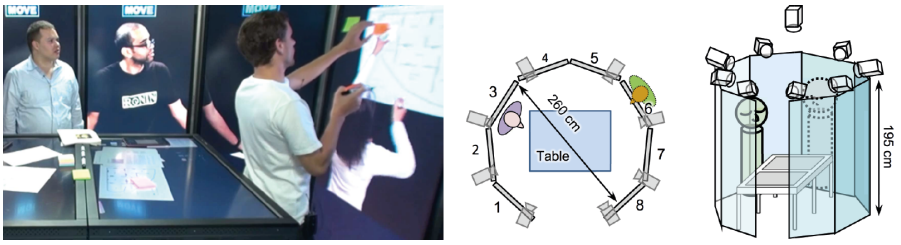


Figure 23: (left) Luff et al.'s *t-Room*, an immersive telepresence system. Image from [78].

One of the few systems that supports existing collaborative practices in large spaces with shared data at a distance is Luff et al.'s [78] *t-Room* (Figure 23). This system is a high-fidelity telepresence system that supports remote collaboration on shared digital objects. The *t-Room* is a space for remote collaboration with a tabletop in the center that contains shared data, surrounded by screens that can display remote participants' video feeds or shared documents. Although the authors do not study this particular setting in a co-located situation to investigate how people use it, they designed the system with the goal of supporting known collaborative practices on large screens and tabletops. The room's layout promotes the use of pointing gestures to indicate far objects, as large screens are sometimes out of reach, and lets people adopt formations around shared content, as there is a tabletop in the center of the room. The authors observe remote collaboration across two *t-Room*s and identify the formations between local and remote participants. They also show that, when remote collaborators look or point at objects, local participants can estimate what is the object being referred to the way in which remote collaboration is realized really takes advantage of the room's characteristics and gives support to existing collaborative practices.

2.4 POSITION OF THIS WORK

The benefits that of having a large high-resolution display in a large space are well established in the literature. Collaboration in particular benefits from such characteristics: people are more aware of each other's actions, they can navigate data both together (tightly-coupled) as well as independently (loosely-coupled), they can arrange information spatially and assign meaning to space, and even manipulate data more efficiently together.

But how can we keep these benefits when collaborators are located remotely? So far, previous work has enabled communication across large interactive spaces by recreating face-to-face situations. From a technological point of view, we can build systems that create a transparent glass illusion between two locations, or even recreate a remote location through a shared 3D virtual world. But in enabling remote *communication* across these spaces, previous work has largely left aside the study of remote *collaboration*. In this dissertation, I study remote collaboration across wall-sized displays to both support existing practices by overcoming the limitations that technology

imposes in remote settings, and enable new possibilities compared to co-located settings.

Previous systems have implemented remote collaboration by going beyond sending simple audio-video signals. With tabletops for example, sending video of a remote collaborator's arm on top of their own screen enables practices that happen normally when co-located, but at a distance, such as switching among coupling styles. Moreover, this technique can create situations that are not possible when co-located, such as changing the way people use the space to create their territories.

I believe that we need to study remote collaboration across wall-sized displays from this perspective, rather than trying to recreate face-to-face situations, as has been done so far. To do so, I start by observing collaborative practices that occur across wall-sized displays, and build a telecommunications system that lets me study how to support them. I explore how video should be captured and displayed to enable existing behaviors during collaboration, such as the use of deictic gestures, and to enable configurations that are not possible when co-located, such as having a face-to-face conversation while standing on opposite ends of the display.

3 OBSERVATIONS

"You can always find abstraction in detail
but you can't find detail in abstraction"

— Wendy E. Mackay

Wall-sized displays bring new opportunities for collaboration. But it's hard to find people that use this technology for collaboration every day. I perform two observations where I use one wall-sized display to simulate two remote locations. In the first observation I use a paper prototype to have a first impression of what remote collaboration might look like. In the second one, I use a technological prototype to capture participants using cameras, and display their video feeds to each other. I find that collaboration depends of two very different types of communication: gestural, when indicating objects on the display, and conversational when holding face-to-face discussions.

Wall-sized displays are hard to come across. Although they have become affordable and have drawn increased researchers' attention over the last decade, it is still difficult to find large high-resolution displays in working environments. This is especially the case for activities that involve collaboration. As a consequence, there are not many users available that master this technology in their daily work practices, limiting the ability to perform real-world observations.

I thus started my exploration by conducting qualitative studies in a simulated remote environment, created by dividing one wall-sized display into two remote locations. I conducted two observations, one with a paper-based prototype and one with a technology-based prototype. Pairs of participants performed collaborative tasks related to their daily work at the time. The goal is to discover what are the practices that a remote collaboration system needs to support for collaboration across these spaces.

3.1 LOW-FIDELITY PROTOTYPE

I devised a low-fidelity paper prototype to have a first impression of what remote collaboration across wall-sized displays might look like, and to understand the requirements of effective collaboration in this setting. I worked with pairs of participants in a Wizard-of-Oz observation that simulates video capabilities across displays.

3.1.1 *System*

I used the *WILD* room for this observation, a large wall-sized display located at the *Paris Saclay University* campus measuring 5m50 ×

Figure 24: Low-fidelity prototype observation: two collaborators working on a shared presentation assembly task; one helper on each side holds an iPad where collaborators can see each other.

1m80. For further details about this room see Appendix A. I divided the room using large cardboard billboards to simulate two remote locations.

On each side, blank sheets of paper, representing blank slides, text clippings and images were laid out on the display. Two helpers held iPads running Skype, such that the participants could see each other through a video link (Figure 24). The participants could hear each other across the room. I simulated shared content by moving back and forth between the two sides and manually syncing their individual changes. The display was turned off and served simply to hold content.

3.1.2 *Participants and Task*

Participants were two females in their early to-mid thirties. They knew each other as they were colleagues at an HCI research lab and had already published together.

Participants had to put together a slideshow presentation based on text and images from a conference presentation that they had recently worked on. They were instructed to create any tool for collaboration that they saw fit, such as defining synchronized elements or spaces, transmitting new text to the remote side or synchronizing object movement. They had no constraints imposed on how to achieve their goal of collaboratively assembling the presentation.

3.1.3 *Data Collection and Analysis*

The session was video recorded and participants were interviewed after they completed the task. Interviews were audio recorded. I watched the videos and listened to the recordings to extract repeated or unusual behavior.

3.1.4 *Results*

During the session, participants generally proceeded in the following way: one side took the lead by starting a discussion on an is-

sue, both reached an agreement on what needed to be done, then the leader made the changes to the presentation while the follower looked. Sometimes the follower also suggested corrections until they reached an agreement.

Oftentimes, collaborators agreed on an object to be synced, mainly a slide, and placed it in the middle. This way, they defined a local private space, and an on-demand shared synchronized space. As they made changes to shared slides by dragging text clippings and images, actions were synchronized by the experimenter moving back and forth between the two sides.

It was interesting to observe that participants often pointed towards objects. When this happened, they were reminded that they could devise a way to somehow transmit this action to the other side although they did not propose other strategies than waving objects in the air and having this movement also be performed on the remote side. The lack of context for pointing made it hard to understand deictic instructions, for instance in this dialogue: *"I was looking at this, and I was thinking.."*, to which the other person responds *"This? What is this?"*.

During the one hour session, participants rarely used each other's video for communication. Their attention was drawn to the objects and the task in front of them. As helpers were holding iPads on the side, turning the body to shift the attention from the task to the remote collaborator was cumbersome. The only times when participants talked while looking at each other's video was when they disagreed, e.g. when trying to decide on the meaning or position of an object.

In summary, the main observations of this preliminary study are:

- participants take a leader-follower approach: one of them guides the other, who follows and expresses her opinion;
- participants rely on deicitc instructions although these are hard to understand, as there are missing context; and,
- participants look at the content laid out on the display much more than to each other's faces, using the iPads to look at each other mostly when they disagreed and discuss the content of an object or of an action to perform on it.

Although insightful, this low-fidelity prototype had several limitations: participants were located in the same physical room and thus could hear each other near, which resulted many times in them facing the cardboard dividing the room while talking, instead of looking at the iPads. Also, they were working while a person held an iPad next to them, and were constantly waiting for a person to reproduce the other side's actions. While participants' behavior was certainly influenced by these artifacts, it was clear that they would look more often at each other's video feed if they were placed on the wall-sized display, in front of them, rather than on the side.

This lead to a second observation based on a prototype that displays participant's faces within the content on the display.

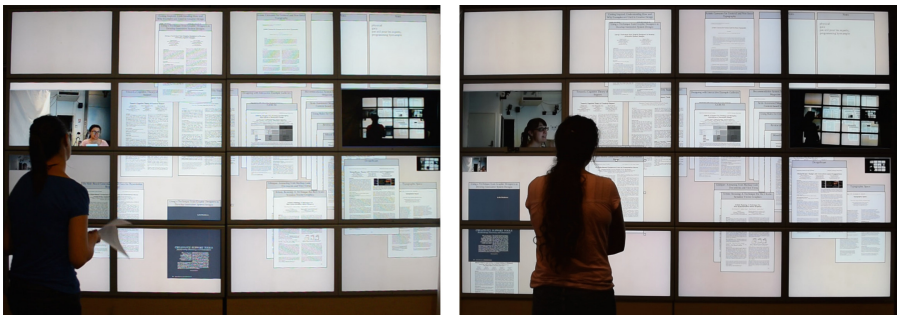


Figure 25: First technology prototype observation: two collaborators working on a literature classification task. Each collaborator has four video feeds: on one of the left-most monitors, the remote person's front camera feed; right below, a smaller feed of their own front camera; on the right-most monitor, the remote person's back camera feed; and, right below, a smaller feed of their own back camera.

3.2 FIRST TECHNOLOGY PROTOTYPE

This prototype was created to capture video of collaborators and display it within the shared content on the remote display. My goal was to observe the use of video feeds during collaboration.

3.2.1 *System*

This prototype consisted of dividing the *WILD* wall-sized display into two parts as in the previous study, this time by hanging a sheet from the ceiling in the middle of the room. I attached two *raspicam*[1] cameras onto the bezels of the left-most monitor of each side, using a custom 3D-printed case and magnets. These cameras are somewhat bulky, but since the bezels of the monitors are quite large, they can be attached without occluding screen pixels. This setup captures users' faces as they stand in front of the display. Each of the two cameras are attached to a *Raspberry Pi 1 model B+*[2] computer mounted on the back of the monitors, connected using a cable slid between two monitors. These computers capture the camera video feed and stream it to a network using *gstreamer*[3].

It became clear form this setup that if participants were out of the camera field of view, they would not be seen by the remote person. To mitigate this, I installed an HD webcam at the back of the room on each side, to capture video of the participants and the wall-sized display from the back as they moved through their part of the display. This video feed was streamed to the network also using *gstreamer*.

Each collaborator had four video feeds on their side of the display (Figure 25): on the second row of the left-most column of monitors, the remote person's front camera feed was displayed, and just below the local front camera feed, in a smaller size. On the second row of

1 https://www.raspberrypi.org/documentation/raspbian/applications/camera.md
2 https://www.raspberrypi.org/products/raspberry-pi-1-model-b/
3 https://gstreamer.freedesktop.org/

the right-most column of monitors, the remote person's back camera feed was displayed and right below a smaller feed of the local back camera. All video streams had a fixed size and position. This made it easy to identify when participants used each video feed, as they moved in front of it. In this prototype, real-time video streams of participants' faces and backs are displayed on top of shared digital content. The experimenter moved content across the screen and turned pages when requested by the participants. *Webstrates* [69] was used to synchronize content.

3.2.2 *Participants and Task*

Participants were two females in their late twenties. They had been working together for 3 months in an HCI lab towards a conference publication. They were asked to sort their related work using a Wizard-of-Oz prototype application built for this session. Papers were laid across the display, their position and current page were synchronized in both sides.

3.2.3 *Results*

Throughout this session, I observed that participants moved for different reasons. First, they physically navigated the space to read the different articles or to see their overall arrangement. Second, they moved to a specific video feed according to the task they were working on:

- they used the front-facing video of the remote person when they wanted to discuss or argue about the content of one particular article or how to cluster it; and,
- they used the back-facing video of the other person when interpreting references to objects and locations, mostly by using deictic instructions (*"this one should go there"*).

This observation showed evidence of two distinct moments in communication: when people need to gesture, and when they need to hold a face-to-face conversation. *Gestural communication* was best supported by the back-facing video, whereas *conversational communication* was best supported by the front-facing video.

In this prototype, communication was fluid and participants were in general satisfied. Nonetheless, they often had to stop what they were doing and move in front of the fixed video feeds to communicate. This interrupted their work and was perceived as annoying. To support communication without disrupting work, it became clear that users' faces need to be captured in front of them, wherever they are standing, and presented on the remote side by taking into account both collaborators movements.

3.3 SUMMARY AND CONTRIBUTIONS

The goal of this thesis is to integrate telecommunication capabilities that allow people to carry out collaborative work across wall-sized displays, supporting the collaborative practices that their activities require. Chapter 2 reviewed previous work that showed how large displays promote physical navigation. This current prototype does not support this practice, as one camera cannot capture all the space in front of the display, and a video feed located in a fixed spot cannot be seen when a person moves away from it. But on the other hand, this limitation allowed me to clearly observe how participants used each camera to fulfill different needs in communication.

I noticed that collaborators need to carry out two very different types of actions during communication: indicating objects on the display and holding face-to-face discussions. These observations are not meant to provide foundation knowledge of how remote communication works, as they are limited in size and scope. But they rather serve as inspiration that guide the direction of the rest of the work in this dissertation. Drawing from these observations, I envision a system that is able to support user movement while enabling both pointing gestures and face-to-face conversation.

4

INITIAL EXPLORATION

This chapter presents a first exploration of how to support user needs during remote collaboration. I perform two experiments. In the first one, I use recordings of actors gazing at and around a camera to determine if perception of direct eye gaze can be conveyed through video in wall-sized displays. In the second one, I use recordings of actors indicating targets on a wall-sized display to determine how accurately people interpret remote pointing gestures. I find that direct gaze is conveyed even when video is located on the display up to 2m away from the observer, and that pointing gestures can be interpreted with great accuracy when video is placed congruent to digital content. Also, that pointing with only the head can be more accurate than pointing with the head and also the arm.

4.1 INTRODUCTION

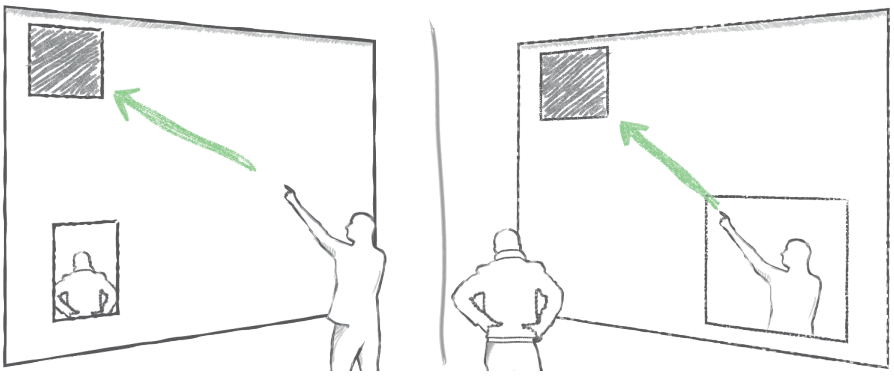


Figure 26: Two collaborators discussing shared data across two wall-sized displays. The person on the left points and gazes at a shared object. The person on the right interprets this action through a video feed.

In the last chapter, I observed that participants engage in two types of communication during remote collaboration across wall-sized displays: *gestural communication* where they rely on the use of deictic gestures to indicate shared objects, and *conversational communication* where they engage in face-to-face talk to discuss about the content of objects.

The prototype used in the last observation did not support remote pointing indications well enough, as participants misinterpreted these actions and had to ask for clarifications. It also limited the way in which participants carried out face-to-face conversation, as they relied on a fixed camera and video feed. It is therefore unlikely that this activity is supported as they move in the space.

In video-mediated communication, remote pointing gesture interpretation depends on the relationship between the person performing the action, the object being referenced, the camera position and the video presented in a remote side. Direct eye gaze interpretation on the other hand, relies on the distance between the recording camera and the spot the person recorded is gazing at. However, it is not clear to what extent the position of the user relative to the video affects the perception of direct gaze and pointing. In this chapter, I investigate the perception of direct eye gaze and pointing gesture interpretation when performed across wall-sized displays.

I recorded two video sets of actors performing actions: in the first one they gaze towards and around the recording camera, in the second one they indicate objects on a wall-sized display using both gaze and their arm. These recordings serve as stimuli during two experiments. In the first one, I investigate (1) how far we can place a camera from a video feed without losing the impression of direct eye-gaze, and (2) whether the position of the observer relative to where the video is displayed affects this perception. In the second experiment, I investigate (1) the accuracy when interpreting remote references to shared objects, performed either by looking or pointing at them, and (2) whether the position of the observer affects this accuracy. This chapter is an extended version of work published at CHI 2015 [5].

4.2 PREVIOUS WORK

4.2.1 *Direct Eye Gaze Perception*

Direct eye gaze perception, or simply eye contact, is when a person has the impression that someone else is looking directly into their eyes. Von Cranach & Ellgring [114] showed that people are capable of interpreting another person's eye gaze with great accuracy. They report that observers as far as 1.5m away and at right angles to the axis between two interactors identified 60% of gaze fixations as being inside the facial region, when these fixations were from one person to the nose bridge of the other. Gibson & Pick [44] reported that participants 2m away from each other correctly judge 84% of gaze fixations to their nose bridge as direct eye contact.

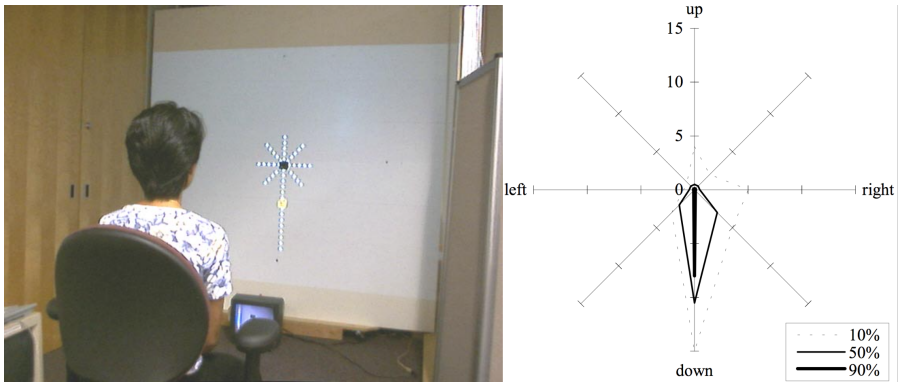


Figure 27: Chen's study of direct eye gaze perception in remote communication. (left) The recording setup. (right) The experiment results. The three curves indicate where eye contact was maintained more than 10%, 50%, and 90% of the time. Images from [24]

This perception is affected by video in technology-mediated communication. Chen [24] studied how far from a camera an observer can gaze before the perception of direct eye gaze is lost. He recorded videos of an actor looking at targets around a recording camera (Figure 27 left), and played this video to participants, asking them to assess if the person was looking directly at them or not. He found that there is still a direct gaze perception 90% of the times up to 5 visual degrees downwards, but only 1 visual degree in other directions (Figure 27 right). This indicates that the best place to put video in a remote conversation and preserve direct eye gaze is below the camera. We can observe from his data that there is a more accurate perception of direct gaze towards the right, which the author does not discuss.

Vertegaal et al. [112] studied groups of three people solving a language puzzle using video-mediated communication. They present remote collaborators varying the degree to which they gaze at the camera. Their results show that when eye contact was not conveyed, participants took about 25% fewer turns. This is because gaze conveys whether a person is being addressed or expected to speak, and is used to regulate social intimacy.

Previous systems have used video to convey direct eye gaze. Dou et al. [34] capture a 3D mesh of people and deform it to correct for eye gaze. Nguyen & Canny's [82] *Multiview* preserves direct eye gaze in a multi-party remote conversation by using multiple cameras to capture people from different positions and present a specific video feed to each collaborator. Vertegaal et al.'s [113] *GAZE-2* is a group video conferencing system that ensures parallax-free transmission of eye contact. They place several cameras behind video to capture people's faces from the spot they are gazing at. They use an eye tracker to select the camera that captures the person from the front, and thus preserves direct eye gaze.

Although direct eye gaze perception has been quantified in face-to-face situations, and some systems have mitigated the effect of video in technology-mediated communication, it is not clear how this perception is altered when the observer is not in front of the video, as it is often the case with wall-sized displays. The fact that in these spaces

people and video of remote collaborators can move, might influence the perception of eye contact.

4.2.2 *Pointing Gestures*

Pointing can be used in conversation to mark an initial reference to an object [80]. Bangerter [9] observed that pointing can provide *joint focus of attention* [28] in spatial regions. Bradel et al. [16] studied pairs performing a sense-making task and observed that pointing actions were used to indicate spatial references to their partners. Pointing can be performed using more than the hands. Griffin and Bock [46] reported that indicating objects can happen also with the eyes and head, as people look at objects while speaking about them even when not pointing explicitly with their hands.

While remote collaboration, the interpretation of remote pointing depends of the relation between the person performing the action and the object being indicated (Buxton's *reference space* [21]). When it cannot be correctly interpreted, communication is hindered. Luff et al. [78] observed breakdowns, when the disparity between gesture production and display created problems when identifying objects during studies across two *t-Room*s.

A number of systems have explored how to combine the *person space* and the *task space* [22] to support pointing in a *reference space*. *VideoDraw* [106] and *ClearBoard* [61] were among the first, by overlying the video feed of the remote collaborator on top of shared drawings on slanted displays. *VideoWhiteboard* [107] also did this integration early on, on a vertical display. More recently, *ConnectBoard* [100] and *Holoport* [73] also realized this overlay on vertical displays.

In large interactive spaces, a number of telecommunication systems have been created that can transmit gaze and pointing. Section 2.3.3 introduced Dou et al.'s [34] Room-sized Informal Telepresence System and Willert et al.'s [115] 2D array of cameras. However, these systems do not consider shared content that participants can point at, as video occupies the whole screen space. Beck et al.'s [11] 3D virtual immersive telepresence supports such content, but it requires a complex technological setup. Neither of these studies featured pointing controlled studies of pointing.

In other contexts however, previous work has indeed measured the accuracy of remote pointing interpretation through experiments. Wong & Gutwin [116] created a Collaborative Virtual Environment (CVE), where users are represented by avatars instead of live video feeds. They found that although pointing gestures are more accurately interpreted in the real world, this difference is small. Akkil et al. [1] studied the accuracy of pointing gesture interpretations in egocentric views. They found that superimposing the gaze information onto the egocentric video can enable viewers to determine pointing targets more accurately and more confidently

In summary, there is great value in accurately interpreting pointing during remote collaboration. Some systems have looked at re-

mote pointing toward shared content, but their setups differ from wall-sized displays. Wong & Gutwin [116] note that determining how accurately viewers can interpret the direction of pointing is one of the fundamental questions to be answered before designing rich support for pointing in collaborative remote systems. Therefore, I set out to study the accuracy with which people can interpret remote indications across wall-sized displays.

4.3 EXPERIMENT 1: PERCEPTION OF DIRECT EYE GAZE

In this experiment, I asses how far to place the video of the remote partner from the camera capturing him, such that there is still direct eye gaze perception, and how the position of the observer relative to the video affects this perception. Participants watch recorded videos of a person looking at targets near the recording camera and answer if they perceive direct eye gaze or not.

The videos show a person looking into 41 targets (8 directions and 5 distances for targets other than the center, and the center target). The videos are shown in a wall-sized display in front of the participants, places at the same position where the recording camera was originally. The video is displayed in 5 different positions on the display.

4.3.1 *Method*

The experiment has a $[5 \times 6 \times 9]$ within-participant design:

- POSITON of the video on the display: *FarLeft*, *Left*, *Center*, *Right* & *FarRight*;
- ANGLE of the gazed targets: *North*, *NorthEast*, *East*, *SouthEast*, *South*, *SouthWest*, *West*, *NorthWest* & *NoAngle* (center target); and,
- DISTANCE of the gazed targets from the center: *Zero* (center target), *One*, *Two*, *Three*, *Four* & *Five*.

POSITON is where the video is presented on the wall-sized display during the experiment, in *Center* the video is in front of the participant, in *FarLeft* and *Left* at 2m and 1m respectively towards the left of the center, *FarRight* and in *Right* at 2m and 1m respectively towards the right of the center. ANGLE is the 8 directions around the recording camera where the actors looked, each at a 45°from each other, and the center. DISTANCE is the 5 distances from the camera where the targets were placed during the recording, and the center (Figure 28 b). Targets in the same ANGLE direction are 1 visual degree apart from each other, when the actor is at a distance of 150cm. For each participant, conditions were grouped by POSITON, which was counterbalanced across participants using Balanced Latin Squares. For each resulting block, the 41 possible ANGLE $\times$ DISTANCE were presented at random.

I use 3 different actors to record the videos, although I do not include it as a factor to avoid having 3 times more the conditions. I cycle

through the list of actor videos so that each participant sees videos of one actor, and all actor are seen an equal number of times.

4.3.2 *Participants*

12 participants (6 male), aged 24 to 38 (mean 29) took part in the study. All participants had normal or corrected to normal vision. 8 had a right (4 left) dominant eye. 1 participant used conferencing systems every day, 6 more than once a week, 3 once a month and 2 almost never. All participants received sweets as compensation for their time.

4.3.3 *Hardware and Software*

The experiment was conducted in *WILD* room (see Appendix A for details on this room). Videos are displayed on the display in 5 positions, at 2225 $\times$ 1252 pixels. Participants answer using an Apple iPad 3 (display: 9.7″), weight: 650g, dimensions: 19 $\times$ 24.3 $\times$ 0.94cm).

4.3.4 *Procedure*

I recorded 41 10-second videos of three different actors looking at targets surrounding a camera: a 29 year-old woman with pulled back hair and brown eyes, a 29 year-old man with brown medium-length hair and brown eyes, and a 27 year-old man with short brown hair and hazel eyes. The camera was placed in front of the actors and they gazed naturally at the 41 targets (Figure 28 a). The actors were sitting at 150mm from the camera and were instructed to look at all targets, one at a time.

Figure 28: Video recording setup. (a) A camera records an actor gazing at the targets (b), printed and placed in front of him. The camera is placed at the center of the targets.

The recorded videos are displayed on wall-sized at participant's eye level while sitting. Based on the focal length of the recording camera (43mm in 35mm equivalent focal length), the size of the video is adjusted so that the remote users appear life-size, as if they are sitting 150cm behind the display—i.e. 300cm away from the participants.

Participants sit 150cm away from the camera. When they are ready to answer, they tap a large "Stop" button on the iPad and choose "yes" or "no" to indicate if they feel direct eye contact (Figure 29).

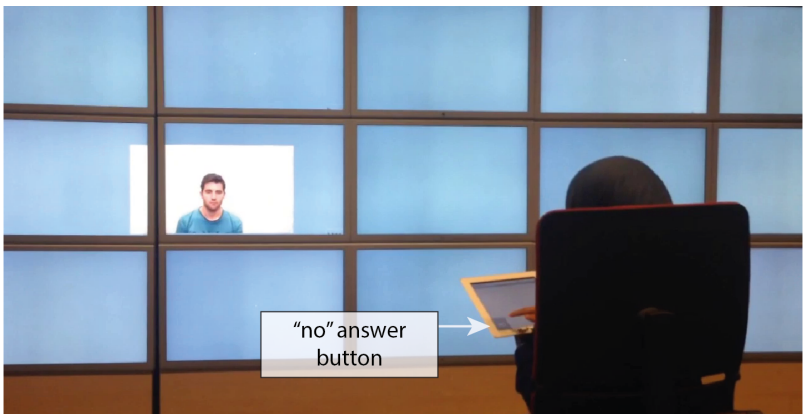


Figure 29: Experiment setup. A participant taking part of the experiment answers "no". The video on the display is at the *Left* position.

During training, participants saw 3 random videos in 4 POSITONS. Then, the 205 videos were presented: 5 POSITON × (8 ANGLE × 5 DISTANCE + 1 target in the center: *NoAngle* and *Zero*).

One trial consists of one video of a target (defined by an ANGLE and DISTANCE) in one POSITON.

4.3.5 *Data Collection*

Participant answer in each trial if they consider that the person in the video is gazing directly at them. I record participant's answers ("yes" or "no"). At the end of the experiment, participants fill out a short questionnaire.

4.3.6 *Data Analysis*

I compute Direct Gaze Perception (*DGP*), to quantify how far from the camera a person can look before the direct gaze perception is lost. To compute the mean value of this measure, I code "yes" answers as 1 and "no" as 0.

4.3.7 *Results*

A total of 2460 trials were registered: 5 POSITON × (8 ANGLE × 5 DISTANCE + 1 target in the center).

4.3.7.1 *Direct Gaze Perception*

As participants' answers are binary, I perform a Wilcoxcon rank sum tests with Bonferroni correction for pairwise comparisons to analyze the Direct Gaze Perception.

Pairwise comparisons for POSITON yield no significant effects on Direct Gaze Perception (all $p > 0.33$) (Figure 30).

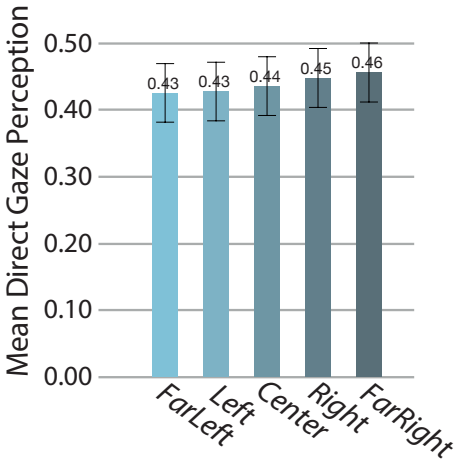


Figure 30: Direct Gaze Perception (*DGP*) by Positon. Bars show 95% confidence intervals.

Pairwise comparisons for Angle show that the following groups are significantly different from each other: {*NoAngle* }, {*North*, *NorthEast*, *East*, *SouthEast*, *South* }, {*SouthWest*, *West*, *NorthWest* } (all p's < 0.0005 in pairwise comparisons) (Figure 31). This shows that there is a more accurate perception of direct gaze for targets on the right, as found by Chen [24].

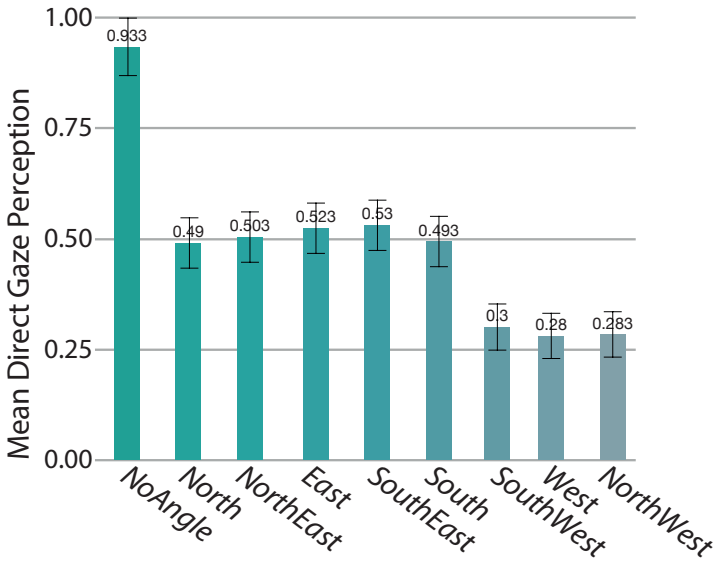


Figure 31: Direct Gaze Perception (*DGP*) by Angle. Bars show 95% confidence intervals.

Pairwise comparisons for Distance show that all levels are significantly different from each other (all p's < 0.0005, except for *Three* vs. *Two* where $p = 0.048$, *Five* vs. *Four* $p = 0.0065$ and *One* vs. *Zero* $p = 0.0015$) (Figure 32).

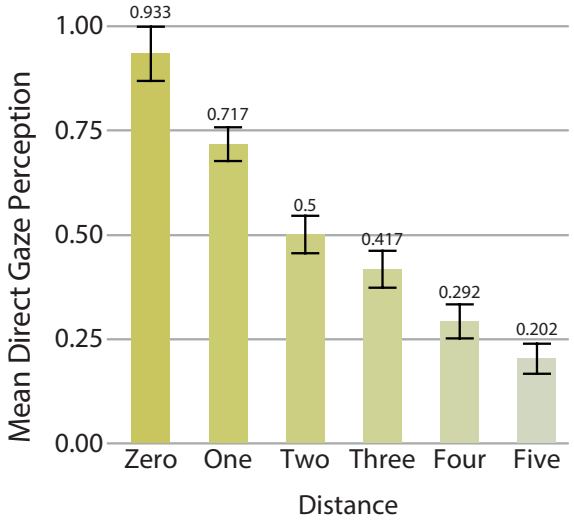


Figure 32: Direct Gaze Perception (*DGP*) by DISTANCE. Bars show 95% confidence intervals.

The plot of mean Direct Gaze Perception by DISTANCE and ANGLE for each POSITON (Figure 33) shows the trend that perception of direct gaze is more accurate for targets on the right than on the left. This confirms Chen [24] results with the added finding that this effect is independent of the viewer's position.

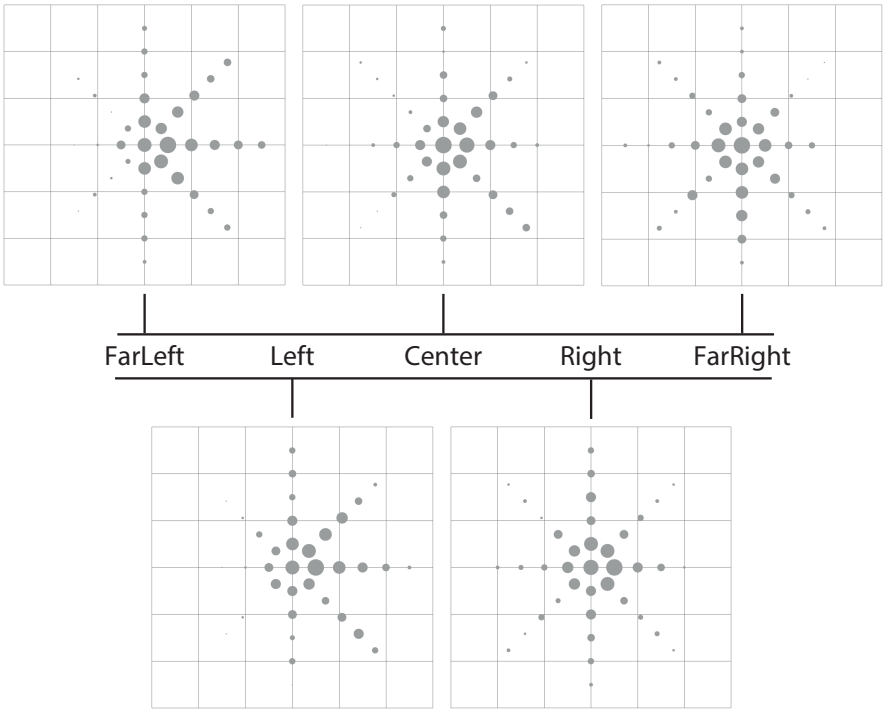


Figure 33: Direct Gaze Perception (*DGP*) by DISTANCE and ANGLE for each POSITON. The size of each dot represents the mean Direct Gaze Perception.

4.3.8 *Discussion*

The most interesting result is on the effect of POSITON. Regardless of where the video was shown, the experiment did not show any effect on Direct Gaze Perception. The results suggest that the *Mona Lisa effect*, where a person on an image seems to be gazing at the observer, is present even when these two are 2m away from each other. This in turn means that, in collaboration across wall-sized displays, video of a remote person does not have to be always in front of a local user, thus gaze is preserved as people move to physically navigate data.

With respect to ANGLE, the result shows a similar effect to that observed by Chen [24], where targets towards the right generate a more accurate perception of direct gaze (Figure 27). ANGLES that are towards the left (*SouthWest*, *NorthWest*, *West*) are different from angles on the center and towards the right (*SouthEast*, *East*, *NorthEast*, *South* and *North*), and both are different from the center (*NoAngle*). Unlike Chen [24] however, we do not observe that targets towards the bottom result in more accurate perception of direct eye gaze.

4.4 EXPERIMENT 2: ACCURACY OF REMOTE INDICATIONS

In this experiment, I assess how accurately an observer can determine which target a remote user indicates on a wall-sized display, and how the position of the observer relative to the video affects this accuracy. Participants watch recorded videos of a person indicating objects on a wall-sized display and answer which object they think the person on the video is pointing at.

19 targets are laid out on the display: one in the center, surrounded by 3 rings of targets along 8 directions, and 2 more targets, one on the far left and one on the far right ($19 = 1 + 8 \times 2 + 2$). I use video recordings to simulate a remote user indicating targets, while participants determine which target is being shown. The videos are presented on the wall-sized display, placed at 5 different positions.

4.4.1 *Method*

The experiment has a $[2 \times 5 \times 4]$ within-participant design, with three primary factors and one secondary factor:

- TECHNIQUE to indicate the targets: *Head* & *Head+Arm*;
- POSITION of the participant in front of the display: *Center*, *FarLeft*, *Left*, *FarRight* & *Right*;
- TARGETDISTANCE where the target is placed: *Ring*0, *Ring*1, *Ring*2 & *Ring*3; and,
- ACTOR (secondary) used for the video recording.

TECHNIQUE corresponds to how actors indicate objects. In the *Head* condition, they use the natural combination of head turning and gaz-

ing, and the *Head+Arm* condition they use the combination of head turning, gazing and pointing with the arm and finger.

POSITION is where participants sit during the experiment. In the *Center* condition they sit in front of the center of the display, in the *FarLeft* and *Left* conditions they sit 2m and 1m respectively to the left of the center, in the *FarRight* and *Right* they sit at 2m and 1m respectively to the right of the center.

TARGETDISTANCE corresponds to the distance from the target to the center. In the *Ring*0 condition the target is on the center behind the video; in the *Ring*1 condition targets lay 11.5 visual degrees apart from the center and from *Ring*2, both rings with 8 targets 45°apart from each other; and in the *Ring*3 condition, 2 targets are 23 visual degrees apart, one on each far end of the display.

ACTOR (secondary) corresponds to the person recorded for the stimuli. 3 actors were used to mitigate the impact that one particular person might have on the results.

For each participant, conditions were grouped by TECHNIQUE, then by ACTOR and then by POSITION. The order in which these were presented was counterbalanced across conditions by using balanced Latin squares for the first three factors. Each Latin square was mirrored and the result was repeated as necessary. In each TECHNIQUE $\times$ ACTOR $\times$ POSITION condition, the presentation order of the 19 targets was randomized so that successive trials never showed targets in adjacent rings from the same direction. When this sequence appeared in pilot studies, it helped participants better estimate the second target based on the first one.

4.4.2 *Participants*

12 right-handed participants (8 male), aged 21 to 33 (mean 27), all computer science graduates, participated in the study. All participants had normal or corrected to normal vision. 10 had a right (2 left) dominant eye. Regarding their background in remote communication, 2 participants used conferencing systems every day, 6 more than once a week, 3 once a week, and 2 almost never. All participants received sweets as compensation for their time.

4.4.3 *Hardware and Software*

The experiment took place in the *WILD* room (see Appendix A for details on this room). Videos are displayed on the center of the display with concentric targets being displayed around it. Participants answer using an Apple iPad 3 (display: 9.7″), weight: 650g, dimensions: 19 $\times$ 24.3 $\times$ 0.94cm).

4.4.4 *Procedure*

I recorded 114 10-second videos of three different actors showing the 19 targets on the *WILD* display for both TECHNIQUES (*Head* and

Head+Arm): a 29 year-old woman with pulled back hair and brown eyes, a 29 year-old man with brown medium-length hair and brown eyes, and a 27 year-old man with short brown hair and hazel eyes. The camera was set in front of the wall-sized display, at the position of the central target during the recordings (Figure 34). The actors were sitting 230cm away from the camera so that the pointing hand was within the recorded frame in all pointing directions. They were instructed to point or look successively at each target.

Figure 34: The setup used for the video recordings. A camera in the center of the targets records an actor indicating them (*Head+Arm* condition in the image).

The recorded videos were displayed on top of the central target, at the same position as the recording camera. Based on the focal length used at recording time (43mm in 35mm-equivalent focal length), the size of the video was adjusted so that the remote users appeared life-size, as if they were sitting 230cm behind the display—i.e. 460cm away from the participants. The height of the chair was adjusted for each participant so that the video was at eye level.

The participants sit in front of the same display used for video recording, 230cm away from the display and watch each video playing in an infinite loop (Figure 35 a). When they are ready to answer, they tap a large "Stop" button on the iPad and choose which target they think the actor is indicating, by tapping it on a replica of the display layout (Figure 35 b & c).

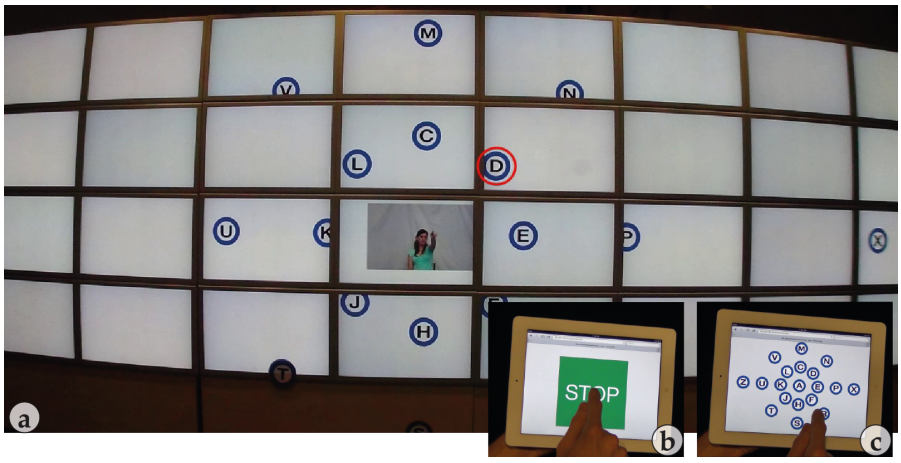


Figure 35: (a) The participant view of the experiment. The actor is pointing to the target "D". The target is highlighted for illustration purposes only, it did not show in the actual experiment. (b) To answer, participants push the "STOP" button on an iPad, and (c) answer by choosing a target.

The 19 targets are displayed on the 5.5m $\times$ 1.8m wall-sized display composed of a grid of 8 $\times$ 4, 30" monitors. Each target is a black letter on a white background surrounded by a blue circle. Letters that could be confused, such as O and Q, are excluded. Targets are distributed in a concentric radial fashion to control both for distance and angle to the video. Three rings of TargetDistance surround a central one (*Ring*0), where the video is displayed (Figure 36). The first two rings (*Ring*1 and *Ring*2) have 8 targets, one for each cardinal and diagonal direction. Due to the aspect ratio of the wall-sized display, the third ring (*Ring*3) has only two targets. *Ring*0, *Ring*1 and *Ring*2 are 11.5°apart when measured from the viewing position (visual degrees), while *Ring*2 and *Ring*3 are 23°apart.

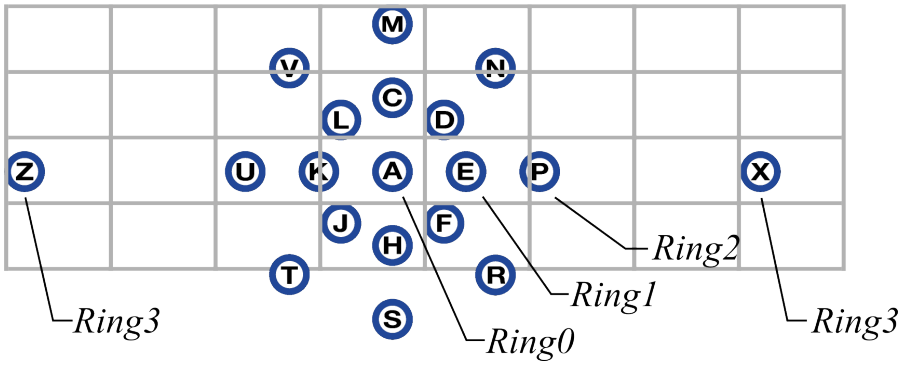


Figure 36: The 19 targets used during the experiment: they lay on a wall-sized display distributed in 4 concentric rings, each with 8 directions.

During training, 4 different random positions from the same subset of 12 videos covering all directions and distances for each participant are used to practice the task and the entry of answers. 4 different random positions are used (2 for the head condition and 2 for the pointing condition) with 3 videos each.

Then, the 570 videos are presented: 2 TECHNIQUES × 3 ACTORS × 5 POSITIONS × 19 TARGETDISTANCES. A mandatory break was imposed every 190 trials (2 sets of 5 POSITIONS), and a reminder of an optional break was provided every 95 trials (5 POSITIONS). Participants could take breaks after seeing the 41 videos in one POSITION.

One trial consists of one video of one ACTOR indicating one TARGETDISTANCE using one TECHNIQUE, while the participant sits in one POSITION.

The experiment lasted about 1 hour in total for each participant.

4.4.5 *Data Collection*

Participants answer in each trial which target they think is being shown. We record the time to provide an answer and the answered target. At the end of the experiment, participants fill out a short questionnaire.

4.4.6 *Data Analysis*

As each target lays at a certain angle and ring from the center, we can compute two measures of error to quantify the accuracy of remote target indication: Distance Error (*DE*) and Angle Error (*AE*). Distance Error corresponds to the difference between the distance of the correct target and the participant's answer. Angle Error is computed as the difference between the angle where both targets lay. Figure 37 depicts an example of how these values are computed.

Trials with the central target as solution are removed when computing *AE*, since these form no angle with other targets.

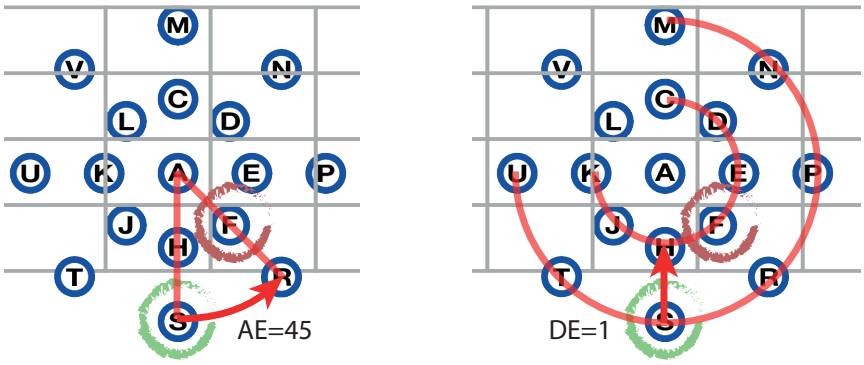


Figure 37: Diagram of Distance Error (*DE*) and Angle Error (*AE*). In this example, the correct target is *S* (circled in green) but the participant answered *F* (circled in red). The left diagram shows the angle between the targets, measured from the center. Here $AE = 45°$. The right diagram shows the circles corresponding to each ring, and the arrow shows the distance error. Here $DE = 1$. For this trial $DE = 1$ and $AE = 45°$.

4.4.7 *Results*

A total of 6840 trials were registered (2 TECHNIQUE × 3 ACTOR × 5 POSITION × 19 targets × 12 PARTICIPANT).

The analysis shows a small learning effect of TECHNIQUE on *DE*: it significantly decreases from 0.36 for the first technique to 0.29 for the second one ($F_{1,6827} = 39.19$, $p < 0.0001$). Many participants learned which videos correspond to the farthest targets and used that information to answer subsequent trials, based on the angle of the head or arm being smaller than that of the farthest targets. There is no evidence of a learning effect on the *AE* ($F_{1,6467} = 3.58$, $p = 0.059$).

In the published article of this work [5], we ran a factorial analysis that tested the 3 main effects with no interactions. In this thesis, I run a full factorial analysis to test for all possible interaction effects. The new effects found with these analysis are marked with a diamond (◆) in the text below.

Unless otherwise specified, full factorial analyses[1] of all measures use the model TECHNIQUE × POSITION × TARGETDISTANCE × Rand(PARTICIPANT) with REsidual Maximum Likelihood (REML) to account for random factors.

4.4.7.1 *Time*

The analysis (◆) of Task Completion Time (*TCT*) yields no significant main effect of TECHNIQUE ($F_{1,6789} = 0.1363$, $p = 0.712$), nor of POSITION ($F_{4,6789} = 1.1133$, $p = 0.3482$), nor of the TECHNIQUE × POSITION interaction ($F_{4,6789} = 0.1695$, $p = 0.954$). It shows a main effect of TARGETDISTANCE ($F_{3,6789} = 7.5190$, $p < 0.0001$), no TECHNIQUE × TARGETDISTANCE interaction effect ($F_{3,6789} = 0.6647$, $p = 0.5737$), nor POSITION × TARGETDISTANCE interaction ($F_{12,6789} = 0.4480$, $p = 0.9442$), and nor TECHNIQUE × POSITION × TARGETDISTANCE interaction ($F_{12,6789} = 0.7076$, $p = 0.7456$). A post-hoc analysis [2] reveals that participants answer faster for *Ring*0 (10.89±4.68s) than all other rings: *Ring*1 (12.88±7.41s), *Ring*2 (12.85±9.01s), *Ring*3 (12.47±7.37s) (all p's < 0.009).

4.4.7.2 *Error*

I report the result of the analysis of each error measure, and then present in detail each effect in a dedicated subsection, as described in Data Analysis. Before presenting the statistical analysis in detail, I highlight that error was very small overall, both for Angle Error (overall mean 0.34±0.52) and Distance Error (overall mean 3.90±12.04).

The analysis of Distance Error (*DE*) yields a significant main effect of TECHNIQUE ($F_{1,6789} = 7.8167$, $p = 0.0052$ ◆), no main effect of POSITION ($F_{4,6789} = 1.0079$, $p = 0.4018$), no TECHNIQUE × POSITION interaction effect ($F_{4,6789} = 0.6089$, $p = 0.6565$). It also shows a main effect of TARGETDISTANCE ($F_{3,6789} = 294.6088$, $p < 0.0001$), a TECHNIQUE ×

1 All analysis are performed using SAS JMP

2 All post-hoc analyses are performed using a Tukey-Kramer "Honestly Significant Difference" (HSD) test

TargetDistance interaction effect ($F_{3,6789} = 12.6101$, $p < 0.0001$ ◆), no Position × TargetDistance interaction effect ($F_{12,6789} = 1.6335$, $p = 0.0753$), and no Technique × Position × TargetDistance interaction effect ($F_{12,6789} = 1.0948$, $p = 0.3593$).

Note that trials with the central target as solution are removed when computing *AE*, since these form no angle with other targets. This difference can be seen in the test's degrees of freedom. The analysis of Angle Error (*AE*) yields a significant main effect of Technique ($F_{1,6439} = 14.2614$, $p = 0.0002$), no effect of Position ($F_{4,6439} = 1.8799$, $p = 0.1110$), no Technique × Position interaction effect ($F_{4,6439} = 0.2624$, $p = 0.9022$). It also shows a main effect of TargetDistance ($F_{2,6439} = 164.7945$, $p < 0.0001$), a Technique × TargetDistance interaction effect ($F_{2,6439} = 8.4239$, $p = 0.0002$ ◆), no Position × TargetDistance interaction effect ($F_{8,6439} = 0.5309$, $p = 0.8341$), and no Technique × Position × TargetDistance interaction effect ($F_{8,6439} = 0.7630$, $p = 0.6356$).

Technique: Post-hoc analyses reveal that for Angle Error, indicating objects with the *Head* leads to smaller errors than pointing with the *Head+Arm* (3.07±12.29°vs. 4.72±15.50°) (Figure 38). But conversely, for Distance Error, indicating objects with the *Head* leads to larger errors than pointing with the *Head+Arm* (0.34±0.52rings vs. 0.32± 0.50rings) (◆).

Surprisingly, using the *Head+Arm* to indicate objects led to higher Angle Error than using only the *Head*. Although significantly different, all errors are relatively small. The effect size for Angle Error is 1.65°and for Distance Error 0.02rings. Although the effect Technique on Distance Error was not previously reported on the publication, its effect size makes this difference rather negligible.

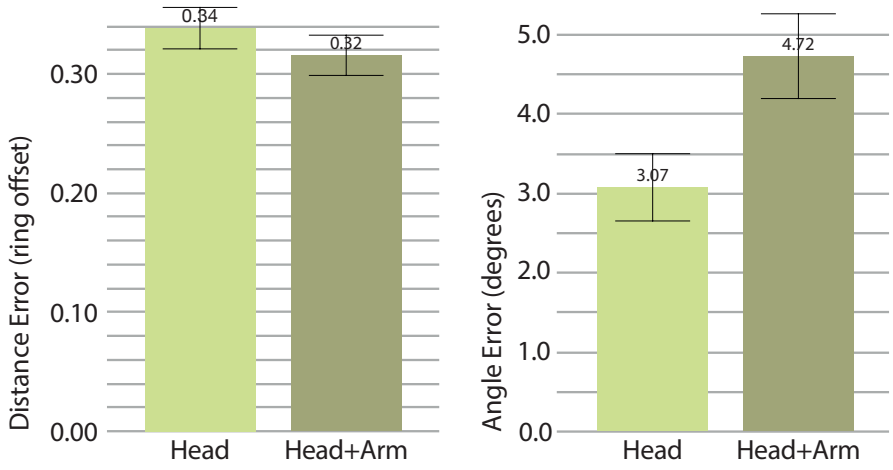


Figure 38: Distance Error (*DE*) and Angle Error (*AE*) by Technique. Bars show 95% confidence intervals.

Position: Positions did not affect Distance Error nor Angle Error (Figure 39).

Distance Error was small in all Positions: *FarLeft* 0.34±0.53rings, *Left* 0.33±0.51rings, *Center* 0.33±0.52rings, *Right* 0.31±0.50rings, *FarRight* 0.32±0.51rings.

Angle Error was also small in all POSITIONS: *FarLeft* 5±15.42°, *Left* 3.47±12.77°, *Center* 3.33±13.05°, *Right* 3.37±12.86°, *FarRight* 4.31±15.63°.

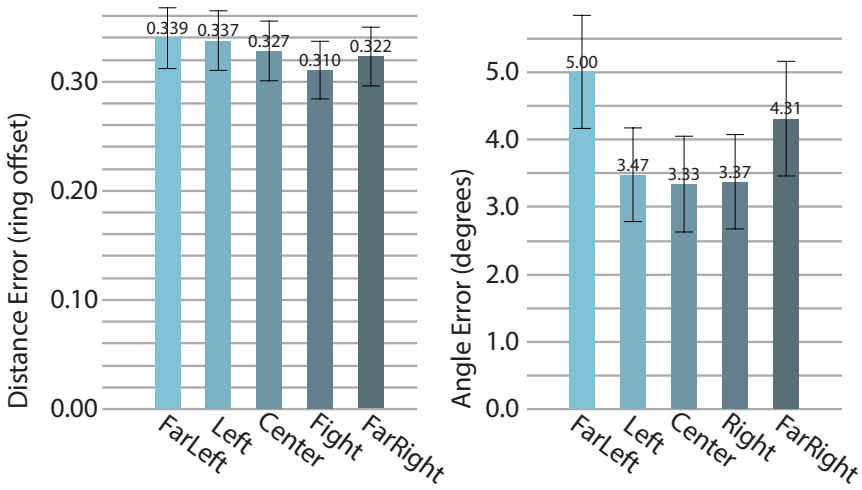


Figure 39: Distance Error (*DE*) and Angle Error (*AE*) by POSITION. Bars show 95% confidence intervals.

TargetDistance: Regarding the main effect, post-hoc analyses reveal that Distance Error increases as the targets are further away, but Angle Error decreases (Figure 40 and Figure 40).

For Distance Error: *Ring*0 (0.18±0.36 rings) and *Ring*1 (0.16±0.37 rings) are significantly different from *Ring*2 (0.44±0.50 rings) which is significantly different than *Ring*3 (0.62±0.78 rings) (all p's < 0.001).

For Angle Error: *Ring*3 (0.75±10.32°) and *Ring*2 (1.28±8.37°) are significantly different from *Ring*1 (7.30±18.01°) which is significantly different than *Ring*0 (13.61±35.68°) (all p's < 0.001).

Regarding the TECHNIQUE × TARGETDISTANCE interaction effect for Distance Error, the following groups are significantly different from each other: {*Head*, *Ring*3 }, {*Head+Arm*, *Ring*3; *Head+Arm*, *Ring*2 }, {*Head+Arm*, *Ring*2; *Head*, *Ring*2 } {*Head*, *Ring*1; *Head+Arm*, *Ring*1; *Head*, *Ring*0; *Head+Arm*, *Ring*0 } (all p's < 0.001 except for *Head+Arm*, *Ring*3 vs. *Head*, *Ring*2 where $p = 0.0394$). What is important among these differences is that for the furthest ring (*Ring*3) the error is higher for *Head* (0.73±0.79 rings) than *Head+Arm* (0.51±0.76 rings) (◆) (Figure 40 right).

Regarding the TECHNIQUE × TARGETDISTANCE interaction effect for Angle Error, the following groups are significantly different from each other:

{*Head+Arm*, *Ring*1 }, {*Head*, *Ring*1 }, {*Head+Arm*, *Ring*2; *Head+Arm*, *Ring*3; *Head*, *Ring*2; *Head*, *Ring*3 } (all p's < 0.001). What is important among these differences is that for *Ring*1 the error is higher for *Head+Arm* (8.88±19.85 rings) than *Head* (5.72±15.82°) (◆) (Figure 40 left).

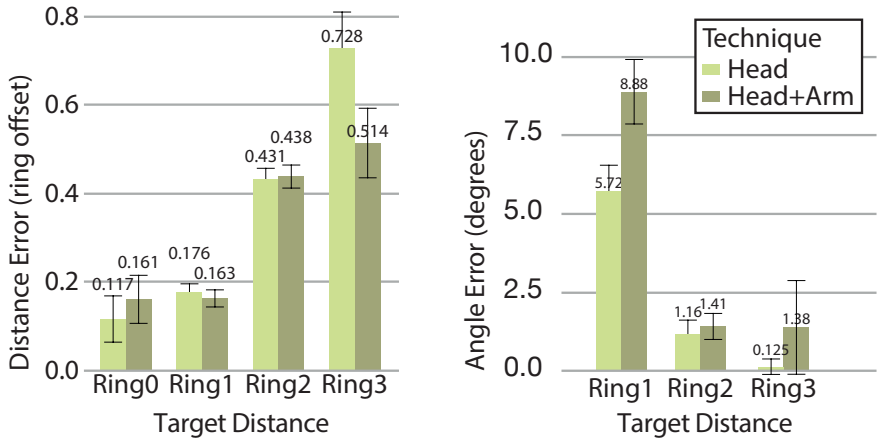


Figure 40: Distance Error (*DE*) and Angle Error (*AE*) by TargetDistance and Technique. Bars show 95% confidence intervals.

4.4.8 *Discussion*

The experiment shows that people can interpret remote pointing rather accurately.

I evaluate error in terms of the distance or angle to the correct target from the video of the remote person performing the gesture. Indicating objects with the *Head* leads to larger Distance Errors (*DE*) than with the *Head+Arm*, although the small effect size makes this difference negligible. Surprisingly, Angle Error (*AE*) is larger when pointing with the *Head+Arm* than with the *Head*. While the effect size is small (1.65°), this was an unexpected result, as one would expect that the arm is the primary source of information to assess the position of the target.

I believe this result comes from the fact that the direction of the arm in the videos does not always indicate the correct target. In natural pointing with the arm, people place the tip of their finger on the line of sight between their eyes and the target [54]. But as video is a 2D representation of the 3D world, it is hard to notice this on a video display. In the post-experiment questionnaire, 4 participants answered that they used only the arm direction when determining the correct target; 6 used first the arm and when in doubt looked at the eyes and head direction; and only one participant connected the eyes with the tip of the finger. This last strategy is actually the correct way to determine the target being pointed. The arm is the most salient cue in the videos when pointing with the head and arm, and it is natural that participants used it as the primary source for determining the correct target, ignoring the geometrical interpretation that we naturally perform in a face-to-face environment.

Participants' position relative to the video on the display had no effect, which is of special importance for remote collaboration across wall-sized display as physical navigation is a common practice in these spaces. This effect is similar to the *Mona Lisa effect*, but for pointing actions instead of gaze.

DE increases as targets are further away from the video. Interestingly, for the outer ring (*Ring*3), the higher error comes from pointing with the *Head*, as with the *Head+Arm* the error is not different from *Ring*2. This result corresponds to the main effect of TECHNIQUE.

AE increases as targets are located closer from the video feed. In this case, the error is largest in the inner ring (*Ring*1) and it is larger for *Head+Arm* than *Head*. This also corresponds to the main effect of TECHNIQUE.

These results on error come from observing a 2D image. When pointing close to the center, a small distance movement, e.g. one ring, produces a very noticeable change in the 2D projection of the arm, but this same change produces small changes when far from the center. Conversely, a small change of angle, e.g. 10°, when pointing near the center results in a very small variation in the 2D projection, but when far away this difference is very noticeable.

Lastly, it is interesting to note that on the farthest targets (*Ring*3), for which it was only possible to express distance, users still made angle errors because they thought that the target laid on *Ring*2.

This experiment shows that there will always be errors as targets are further away from the video of the person indicating them, and systems should take this trade-off into account when placing video. Data shows that while distance error increases with distance to the video, angle error decreases.

4.5 IMPLICATIONS FOR DESIGN

I derive the following implications for the design of telecommunication systems for collaboration across wall-sized displays.

4.5.1 *Telepointers Are Not the Only Solution to Indicate Objects*

This is so because people can accurately estimate the position of a target by looking at the hand and head of a remote person indicating an object. Although hand pointing has been implemented in early systems, such as *VideoDraw* [106], *VideoWhiteboard* [107] or the *t-Room* [78], in these systems the hand actually makes contact with the object being pointed. In this study, I show that collaborators can interpret remote pointing gestures when these are performed at 1.5m from the target, as long as the video of the remote pointing gesture is placed congruent to the digital shared content.

4.5.2 *Hands Are Not Always Needed to Indicate Remote Objects*

Collaborators can indicate objects with great accuracy by using only the head. This means that people can use deictic instructions while e.g. holding tools in their hands to interact with content.

4.5.3 *Video Feeds of Remote Collaborators Can Be Moved Across a Wall-Sized Display*

We can move video feeds of remote collaborators as their position does not greatly affect the interpretation of remote indications. This supports remote pointing gestures across the whole display surface as we can match the video position on one display with the remote camera's position on the other one, and thus keep the correct spatial relationship between shared content and video of collaborators.

4.5.4 *Collaborators Can Move in Front of a Wall-Sized Display*

Collaborators can move without hindering the accuracy of their interpretation of remote indications. This effect is analogous to the *Mona Lisa effect* but for pointing gestures. One possible solution for capturing video can thus be to use multiple cameras that can follow people as they move in front of the display, e.g. to explore the content, similarly to the camera arrangement proposed by Willert et al. [115]. Presenting the video on the remote display, however, needs to be further studied as it plays an important role in communication.

4.6 SUMMARY AND CONTRIBUTIONS

In this chapter I investigated the perception and accuracy of direct gaze and of remote object indication across two wall-sized displays.

Regarding direct gaze, the results on the combination of target distance and angle to the center are hard to interpret, so it is not possible to make solid claims about where the camera should be placed with respect to video and convey direct eye gaze. Data shows that looking towards the right of the camera tends to generate more direct gaze perception than in the other directions, but this effect is not as clear as in previous work [24]. However, the experiment shows that the relative position between the video and the observer does not seem to affect direct gaze perception (up to 2m), suggesting that the *Mona Lisa effect* is present as people move in front of the display.

Regarding remote pointing, the experiment shows that placing video congruent to digital content in a wall-sized display creates a *reference space* [21], where pointing gestures can be interpreted. The results show that error of interpretation is generally small, suggesting that pointing gestures can be used in remote collaboration without hindering the communication process. Producing pointing gestures only with the head leads to smaller angle errors than using the arm, which means that pointing gestures can be perceived even when they are not performed explicitly. Lastly, the observer's position has little to no effect on accuracy, suggesting that we can move the remote person's video feed without disrupting collaboration. This effect is analogous to the *Mona Lisa effect* which is here extended for pointing.

The size and resolution of wall-sized displays leads collaborators to point towards objects and physically move while exploring data.

In the next chapter I use these results to design a telecommunications tool that conveys direct eye contact and pointing gestures across wall-sized displays.

5

CAMRAY: CAMERA ARRAYS FOR REMOTE COLLABORATION

This chapter presents CamRay, a telecommunication system that uses an array of cameras mounted on tiled displays to capture people's faces as they move, and presents their video on top of existing content on remote walls. I present the system's implementation, including deployment differences between WILD and WILDER. I also present performance tests, where I evaluate CamRay in one room to leave out possible network delays, and then across two locations to take into account the delays. CamRay can be used to perform observations in a wide range of scenarios. It can be used to discover the needs for collaboration across large interactive spaces and to support them.

5.1 INTRODUCTION

I observed in Chapter 3 that when collaborators work across wall-sized displays, (1) they make a notable use of pointing gestures towards digital objects and (2) they hold face-to-face discussions when talking about the content of such objects. My goal is to develop a telecommunication system that satisfies these needs through capturing video of remote collaborators and displaying it in a local remote wall-sized display. For this I build on last chapter's findings that we can present video of a remote collaborator on a wall-sized display far from the local user (up to 2m) and still have direct eye gaze perception and a high accuracy when interpreting pointing gestures.

In this chapter, I present *CamRay*, a system that uses camera arrays mounted on tiled displays to capture people's faces as they move, and presents their video on top of existing content on remote walls. *CamRay* is based on open technologies and is designed with a pipeline that allows each component to be executed independently. Its distributed architecture allows for high flexibility and interchangeability of parts, which makes it scalable in the size of the displays, the number of users at each site and the number of sites. New cameras that can capture people from different angles can be easily added, as well as displays that can present video feeds independently from the wall sized-display.

5.2 TELECOMMUNICATION ACROSS LARGE INTERACTIVE SPACES

Remote communication across large interactive spaces has been previously realized using different approaches, as discussed in Section 2.3.3. One way is to increase the sense of telepresence by reproducing remote sites as faithfully as possible. Beck et al.'s [11] immersive group-

to-group telepresence system uses a cluster of depth cameras to capture people, and presents them using a realistic 3D reconstruction in an immersive shared virtual world. This system supports multiple users at each site by using perspectively correct stereoscopic images. Maimone et al. [79] also use multiple depth cameras to capture users, this time with a focus on creating a clear 3D reconstruction of the remote person, filling its holes and smoothing it. Although these systems are impressive from a technological perspective, the authors did not explore how they can solve the needs of effective collaboration.

Another way in which remote collaboration across wall-sized displays has been realized is to create an extended window metaphor—the illusion of a transparent glass between two sites. Willert et al. [115] use a 2D array of cameras mounted on the tiles of a wall-sized display to capture video and display it remotely with a parallax effect to create the transparent glass from the metaphor. Dou et al. [34] also implement the extended window glass metaphor and improve direct gaze perception. These systems allow for groups to communicate at a distance, but they do not consider collaboration using shared digital content.

I draw inspiration from this previous work and design a telecommunications system that connects two remote wall-sized displays and supports remote collaboration across them.

5.3 CAMRAY

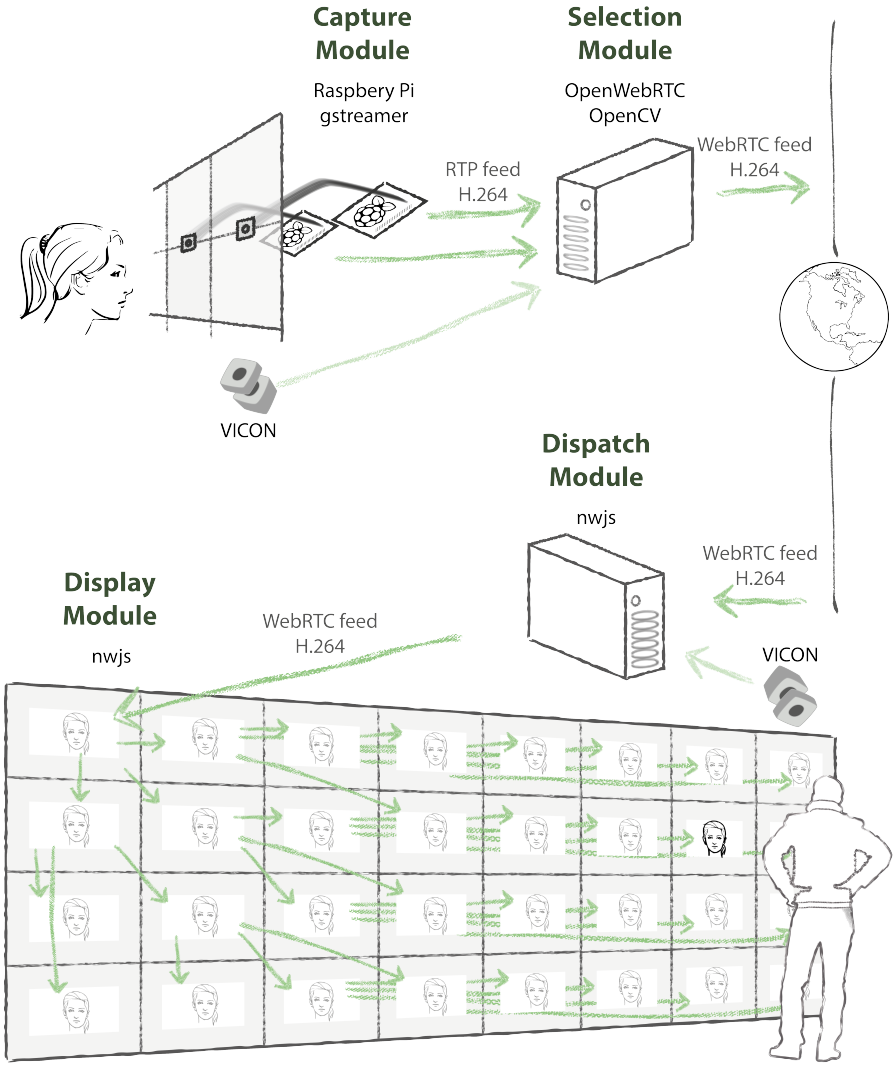


Figure 41: *CamRay* architecture. *Capture Module*: a Raspberry Pi with a camera. *Selection Module*: receives RTP H.264 video feeds, selects one using position data from the *VICON* motion tracking system, and forwards a *WebRTC* stream to a remote site. *Dispatch Module*: receives a *WebRTC* video stream and dispatches it to one *Display Module*; also controls the position of all *Display Module* based on *VICON* data. *Display Module*: receives a *WebRTC* video stream, presents it, and forwards it to other *Display Module*s.

I created a telecommunications system that takes into account physical navigation for capturing video of users' faces and leverages this movement when presenting video feeds on remote displays. Capturing users' faces as they move in the space requires more than one camera, so I used an array of cameras to cover user movement along the display. These cameras are embedded on the bezels of the display tiles, to capture users' faces from the front as they work with data on the display. I chose to work with RGB cameras, as opposed to depth sensors, as this is sufficient for the type of interaction I am

interested in. My goal is to display users' upper bodies and faces during collaboration, not to segment and differentiate the users from the background as Dou et al. [34]. Also, I want to enable collaboration with shared content on wall-sized displays, not recreate a 3D virtual environment to work with 3D data as Beck et al. [11].

I call this system *CamRay*, as its most salient component is an array of cameras on top of existing wall-sized displays. *CamRay* has a flexible infrastructure that allows to add cameras and display video feeds in flexible ways. This allows me to create different combinations of video capture and display to study how to support remote collaboration across wall-sized displays. The system is composed of several modules:

- *Capture Module*: captures video;
- *Selection Module*: selects the video stream from the camera in front of a person;
- *Dispatch Module*: forwards a video stream to a client on the wall-sized display and controls the position of all clients; and,
- *Display Module*: receives a video stream, displays it and optionally forwards it.

In the next sections I detail each module, including the technologies used in the current implementation. I also explain how the *Capture Module* is implemented in both *WILD* and *WILDER*, as the bezels in the monitors of each space are different.

5.3.1 *Capture Module*

This module is composed by a camera and a computer that forwards an RTP video stream in H.264 format on the a network. The camera is mounted on the lower bezel of a monitor from the display, and the cable that connects it to the computer is slid between two adjacent monitors. This allows to mount the cameras to the displays without occluding content. A Raspbery Pi[1] computer with one of the available cameras built for this computer provides good video quality and enough speed to send video to the network in real time.

To capture video from the camera, the Raspbery Pi includes the command *raspivid*[2]. To forward video to the network in RTP format, I use the tool *gstreamer*[3], an open-source library for streaming video.

The following command captures and forwards video:

```
$ raspivid  --verbose --nopreview -hf -w 800 -h 600
  --framerate 30 --bitrate 1000000 --profile baseline
  --timeout 0 -o - | gst-launch-1.0 -v fdsrc ! h264parse !
  rtph264pay config-interval=1 pt=96 ! udpsink
  host=<ip> port=<port>
```

1 https://www.raspberrypi.org
2 https://www.raspberrypi.org/documentation/usage/camera/raspicam/raspivid.md
3 https://gstreamer.freedesktop.org/

The `raspivid` options specifies how to capture video as follows:

- `- verbose` outputs debugging and information messages during the program run, which can be safely omitted;
- `-hf` flips the image horizontally;
- `-w 800 -h 600` sets the size to 800 × 600;
- `-framerate 30 -bitrate 1000000` sets the framerate and bitrate of the video (10Mbits/s here);
- `-profile baseline` sets the H.264 profile used for the encoding;
- `-timeout 0` sets no time limit to terminate the command; and,
- `-o -` outputs the result into the terminal, allowing it to be piped to the command `gst-launch-1.0`.

The `gst-launch` options specifies the pipeline used to process video as follows:

- `-v fdsrc` reads data from a Unix file descriptor;
- `h264parse` parses a H.264 stream;
- `rtph264pay config-interval=1 pt=96` encode H.264 video into RTP packets; and,
- `udpsink host=<ip> port=<port>` sends UDP packets to the specified host and port.

The output of the module is a H.264 stream over UDP, sent to a specified host in the local network.

The next sections provide details on the specific implementation of the *Capture Module* in each wall-sized display.

5.3.1.1 *Capture Module in WILD*

Figure 42: The *CamRay* Hardware in the *WILD* wall-sized display. (a) The *WILD* wall-sized display with one camera in each column on the top row of monitors. (b) A closeup of the *Camera Module for Raspberry Pi*. (c) The *Raspberry Pi* computer at the back of a monitor with its camera attached.

The *WILD* room has a high-resolution wall-sized display composed of 32 monitors in a 4 × 8 grid. Its total resolution is 20480 × 6400 = 131,072,000 pixels. It is controlled by a cluster of 163.2GHz quad core Apple Mac Pro with 2 Nvidia 8800GT graphics cards each, located in a dedicated server room. The software that runs on the monitors can be controlled by three computers in the room, called *frontal*1—3. This room is equipped with a *VICON* infrared real-time motion tracking system with an accuracy of 0.5mm. For more details on the existing infrastructure of this room see Appendix A.

There are 8 *Capture Module*s in the *WILD* room. Each module is composed of a *Camera Module for Raspberry Pi*[4] that measure 25mm × 24mm. The cameras are mounted onto the bezels of each top-row monitor using custom 3D printed mounts held on the bezels by magnets (Figure 42). The cameras do not occlude content as they are placed on the bezels, which are fairly wide. They are placed at 194cm from the ground with a 27° angle. These values were obtained through testing both in the first and second row of monitors at several angles, while checking that peoples' faces were captured both when close—touching objects on the display—and when far back—up to about 1.5m. Each camera is connected through a flat cable that slides between two monitors to a *Raspberry Pi 1 model B+*[5], which is attached to the back of the monitor where the camera is mounted.

The output of each *Capture Module* is sent to a front-end computer *frontal*3 through a dedicated network, where the *Selection Module* is running.

5.3.1.2 *Capture Module in WILDER*

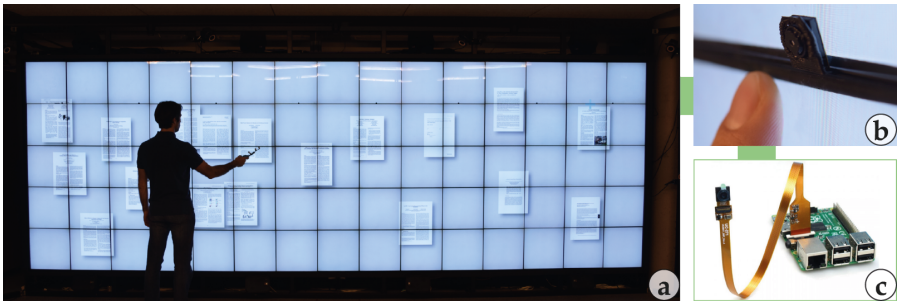


Figure 43: The *CamRay* Hardware in the *WILDER* wall-sized display. (a) The *WILDER* wall-sized display with one camera in each column on the top row of monitors. (b) A closeup of the *Spy Camera for Raspberry Pi* module. (c) The *Raspberry Pi* computer at the back of a monitor with its camera attached.

The *WILDER* room has a high-resolution wall-sized display composed of 75 monitors arranged in a 5 × 15 grid. Its total resolution is 14400 × 4800 = 69,120,000 pixels. It is controlled by a cluster of 103.7GHz quad core Intel Xeon PCs with an NVIDIA Quadro K5000 graphics card each, located in a dedicated server room and connected to the monitors through fiber optics. There are three computers that run software for the display, called *master*1–3. This room is also equipped with a *VICON* infrared real-time motion tracking system. For details on the existing infrastructure of this room see Appendix A.

There are 8 *Capture Module*s in the *WILDER* room. Cameras in this room need to be much smaller in order to occlude as little content as possible, as the bezels are very thin (<1mm). Each module is composed of a *Spy Camera for Raspberry Pi*[6] that measures 8.5mm ×

4 https://www.raspberrypi.org/documentation/hardware/camera/
5 https://www.raspberrypi.org/products/raspberry-pi-1-model-b/
6 https://www.adafruit.com/products/1937

11.3mm. Cameras are mounted on the lower part of the top row of monitors using custom 3D printed mounts (Figure 43). They are spaced out in the same fashion as in *WILD*, such that the content displayed next to one camera is the same in both rooms. Custom 3D printed mounts hold cameras between two monitors at 1m84cm from the ground at and angle of 22°. They also hold the cable that goes towards the back of the monitor and connects the camera to a *Raspberry Pi 2*[7].

The output of each *Capture Module* is sent to a front-end computer (*master*3) through a dedicated network, where the *Selection Module* is running.

5.3.2 *Selection Module*

This module receives the video streams from each *Capture Module* and a data stream from the *VICON* motion capture system containing the user position. In this implementation, I use the *WILD Input Server* [85] to capture the *VICON* data and forward each person's position in space to the *Selection Module*. The advantage of doing so is that the *WILD Input Server* already computes these coordinates in each rooms' space.

The module is implemented using the C++ *OpenWebRTC*[8] implementation. I modified the source code of this framework to receive RTP streams over the network using *OpenCV*[9] compiled with *gstreamer* support.

To read OSC messages from the *WILD Input Server*, the current implementation uses the *liboscpack* package [10]. This can be easily changed to any other protocol (e.g. WebSockets) for receiving user position data, depending on how this data is sent.

The code snippets below show the modifications to the *OpenWebRTC* implementation, and are useful for building future implementations of *CamRay*. These snippets work for 8 cameras.

The `main` method reads all the RTP streams. Then it computes the width and height of the output stream based on the first frame received from the first stream, in order to use the same resolution for output as for input. Finally, it sets the `url` variable with the correct IP, port and session id of the *Dispatch Module* where the stream will be sent. The session id must be the same in both modules, as required by the *WebRTC* protocol.

```
1 gint main(gint argc, gchar **argv)
2 {
3   // Read from all camera feeds
4   static int ports[] = {9000, 9001, 9002, 9003, 9004,
        9005, 9006, 9007};
5   static const int numberOfCameras = 8;
6
```

7 https://www.raspberrypi.org/products/raspberry-pi-2-model-b/
8 http://www.openwebrtc.org/
9 http://opencv.org/
10 https://packages.debian.org/sid/liboscpack-dev

```
7    static VideoCapture video[numberOfCameras];
8
9    for (int i = 0; i < numberOfCameras; i++) {
10     String cmd = "-v udpsrc port=" +
           std::to_string(ports[i]) + "
           caps=\"application/x-rtp\\,\\
           media\\=\\(string\\)video\\,\\
           clock-rate\\=\\(int\\)90000\\,\\
           encoding-name\\=\\(string\\)H264\" !
           rtph264depay ! avdec_h264 ! videoconvert !
           appsink sync=false";
11
12     video[i] = VideoCapture(cmd);
13   }
14
15   // Get width and height from the first frame
16   Mat frame0;
17   video[0] >> frame0;
18   width = frame0.cols;
19   height = frame0.rows;
20   nb_channels = frame0.channels();
21
22   // Gest session ID from arguments
23   session_id = argv[1];
24
25   // Connect and launch the OpenWebRTC connection
26   gchar *url;
27   client_id = g_random_int();
28   serverURL = "http://<ip>:<port>";
29   url = g_strdup_printf("%s/stoc/%s/%u", serverURL,
       session_id, client_id);
30
31   owr_init(NULL);
32   owr_get_capture_sources(OWR_MEDIA_TYPE_VIDEO,
       (got_local_sources, url);
33   owr_run(); // Run main loop
34 }
```

A callback to a custom method called `fill_data` is attached inside the `handle_offer` method from OpenWebRTC.

```
1  OwrAppMediaSource *app_source =
       owr_app_media_source_new("",
2    OWR_MEDIA_TYPE_VIDEO,
3    OWR_SOURCE_TYPE_CAPTURE,
4    "BGR", width, height,
5    nb_channels, (gint) framerate,
6    G_CALLBACK(fill_data));
7  source = OWR_MEDIA_SOURCE(app_source);
8
9  owr_media_session_set_send_source(media_session,
       source);
```

`fill_data` provides provides the frames to be sent to the remote location. The variable `facing_screen` must be filled with the camera number that the user is in front of.

```
/* Adding to fill the data of the appsrc */
static gpointer fill_data()
{
  // Result feed to return, according to user selection
  Mat* result = nullptr;

  // Default case: take the camera that the VICON
  video[facing_screen] >> frame[facing_screen];
  result = &frame[facing_screen];

  // Default
  if(result == nullptr){
    result = &frame[0];
  }

  return result->data;
}
```

Note that the following global variables have to be defined:

```
static gint width;
static gint height
static const gchar *serverURL;
int facing_screen = 0;
```

Currently, this module runs on the front-end computers of *WILD* (*frontal3*) and *WILDER* (*master3*). The output of the module is a H.264 stream using the *WebRTC* protocol, sent to a remote host over.

5.3.3 *Dispatch Module*

This module is implemented in *nwjs*[11] (previously known as node webkit), based on Ericsson Research's *channel server*[12]. This server is an OpenWebRTC implementation that can be called from web clients.

The module uses *peerjs*[13] to connect all the web clients that send and receive streams. A local version of *peerjs* can be used, or the cloud service provided by the framework developers.

I provide code snippets to implement the web client that composes this module, based on Ericsson Research's demo application[14].

In the page loaded event, the page listens to remote *WebRTC* connections:

```
// Join Peers automatically
var sessionId = <same session id as Selection Module>;
signalingChannel = new SignalingChannel(sessionId);

// A peer has joined our session
signalingChannel.onpeer = function (evt) {
  peer = evt.peer;
  peer.onmessage = handleMessage;

```

11 http://nwjs.io/
12 https://github.com/EricssonResearch/openwebrtc
13 http://peerjs.com/
14 https://demo.openwebrtc.org/

```
10    // When a peer has joined, start the call
          automatically
11    start(true);
12  };
```

The methods `start` and `handleMessage` in the demo application handle the sdp handshake and initiate the connection.

I modified the `onaddstream` callback within the `start` method to forward the video stream as soon as it is received.

```
1   function start(isInitiator) {
2     pc = new RTCPeerConnection(configuration);
3
4     ...
5
6     pc.onaddstream = function (evt) {
7       remoteView.src = URL.createObjectURL(evt.stream);
8
9       // Call First Wall-Sized display Display Module
10      peerJS.call(settings.platform.firstPeerToCall,
            evt.stream);
11    };
12
13    ...
14  }
```

This module also receives user position data, so that it can control the position of all *Display Module*s running on the wall-sized display. The current implementation uses WebSockets[15]to listen to *WILD Input Server* messages. The details of this process depend on the dimensions and resolution of each display, so currently two JSON objects define the characteristics of *WILD* and *WILDER*. When this module starts, a parameter defines which configuration to use. Then, the module connects to every *Display Module* specified using a *peerjs* connection, in order to send messages with position updates.

Currently, this module runs on the front-end computer of *WILD* (*frontal3*) and in *WILDER* (*master3*). The output of the module is a H.264 stream using the *WebRTC* protocol, sent to a specified host in the local network.

5.3.4 *Display Module*

This module is implemented in *nwjs*. It is a simple web client that receives a *WebRTC* video feed and displays it, and accepts *peerjs* connections to receive messages with position data.

In a first attempt, I tried to send video only to one monitor and its neighbors, to account for the case when the feed must span more than one display. However, the cost of opening and closing connections ruled out this strategy: in several tests, walking fast in front of the display led to black video feeds while the connection was being established.

15 https://tools.ietf.org/html/rfc6455

I decided then to send the video feeds to all *Display Module*s and only show the one that needed to be seen. Currently, one *Display Module* receives the *WebRTC* video stream from the *Dispatch Module* and forwards it using a tree structure to all other clients on the wall-sized display.

This tree-shaped distribution is designed to balance the load of each node. I tried to make the *Dispatch Module* send the video stream to all clients, but *peerjs* could only handle a small number of connections, thus only around 16 clients actually received the video. As I aim at a scalable architecture, this option is not adequate, especially for *WILD* where video needs to be sent to 32 clients. A daisy-chain distribution, where each client forwards the video to the next one, is also not viable as each *Display Module* adds a small delay, which becomes. I thus decided to use a tree-shaped distribution, where each client forwards the video to 3 other clients (Figure 41). This strategy connects all the clients of the tiled display with a small number of video forwards by each client. The result is an imperceptible delay between each client, allowing to have video in all *Display Module*s with low latency.

Because the *Dispatch Module* controls the position of each video, it can hide all *Display Module*s except for one. As the tree-shaped distribution makes delays imperceptible, a video feed that spans more than one monitor shows no tearing of the image.

5.3.5 *Performance Tests*

I carried out two performance tests to evaluate the performance of *CamRay*. In the first one, I measure the delay of the system in *WILDER* only, by looping the video from the camera array back to the local display. This allows to measure the delay of the system, without adding possible network delay. In the second test I evaluate the performance across *WILD* and *WILDER*, to account for the network delay.

All tests were performed with video captured at 3 different resolutions: 800×600, 1280×960 which are of 3:4 format, and 1920×1080 which is the maximum resolution available in the cameras.

The tests included the following computers, with the following specifications:

- *frontal*3 in *WILD*, a Mac Pro (early 2009), with $2 \times$ 2.26GHz Quad-Core Intel Xeon processors, 8GB 1066MHz DDR3 ECC RAM memory and an NVIDIA GeForce GT 120 512MB graphics card; and,
- *master*3 in *WILDER*, a iMac Retina 5k (27-inch, late 2015), with a 3.2GHz Intel Core i5 processor, 16GB 1867MHz DDR3 memory and an AMD Radeon R9 M390 2048MB graphics card.

The *Capture Module*s used for the test in *WILD* and *WILDER* are the same as described in Section 5.3.1.1 and Section 5.3.1.2.

Figure 44: Setup for *CamRay*'s performance tests. (a) A timer was recorded by one of the *Capture Module* cameras mounted on *WILDER* and looped back to the display for the local performance test. Both the timer and it's video are recorded by a video camera. (b) For the remote performance test, a mirror reflected the video of the timer in *WILD*, which was recorded by a *Capture Module* in this room and sent back to *WILDER* for the remote performance test.

5.3.5.1 *Local Performance Test*

The goal with this test is to measure the delay of the system without accounting for network delays. I thus run the test in one room only (*WILDER*), by using *CamRay* to record video of a running timer and display it in the same display (Figure 44 a), without relying on external networks. I use an external camera to record the live timer and its video feed after being processed by the system. In this way, the original time and the time after capturing, processing and displaying the video are visible at the same time. After the test was finished, I used the video recording to sample the delay every 10 seconds. The delay was computed as the difference between the value on the timer and the value on its video feed.

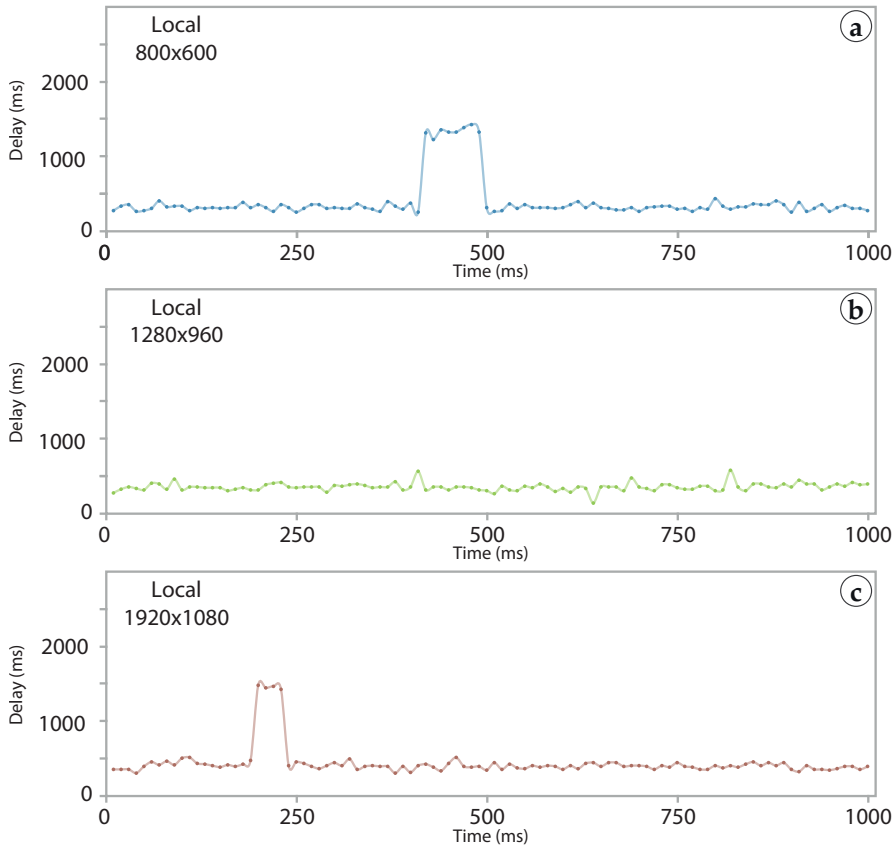


Figure 45: *CamRay* local performance test results with 3 different resolutions: (a) 800 × 600, (b) 1280 × 960 and (c) 1920 × 1080.

Both the *Selection Module* and *Dispatch Module* were running on *master*3.

For a resolution of 800 × 600 (Figure 45 a), the mean delay is 396± 280ms, with a median of 310ms. The delay averages 310ms, and for a brief moment it goes up to around 1400ms. For a resolution of 1280 × 960 (Figure 45 b), the mean delay is 350±54ms with a median of 350ms. For a resolution of 1920 × 1080 (Figure 45 c), the mean delay is 441±214ms with a median of 400ms. Again in this case there was a brief moment where the delay went up to 1400ms.

In general the delay is acceptable for remote communication. There were brief moments in the first and third tests where the delay was higher. This can be attributed to the process scheduling of the computer where the *Selection Module* and *Dispatch Module* are running, or to delays in the network of *WILDER*'s cluster, which connects the *Dispatch Module* and the *Display Module*s.

5.3.5.2 *Remote Performance Test*

This test uses two rooms (*WILD* and *WILDER*), and is intended to measure the end-to-end delay of the system including network delays. I captured video of a timer using *CamRay* in *WILDER*, then using a mirror the system re-captured the output video in *WILD* and looped it back to *WILDER*, where a camera filmed both the timer and the video after it had been sent back and forth between the rooms (Fig-

ure 44 b). This way, the original time and the time after the two-way trip were visible at the same time. I sampled every 10 seconds after the test, and computed the two-way delay based on the two timers. I divided this delay by 2, to compute the delay of sending and receiving video.

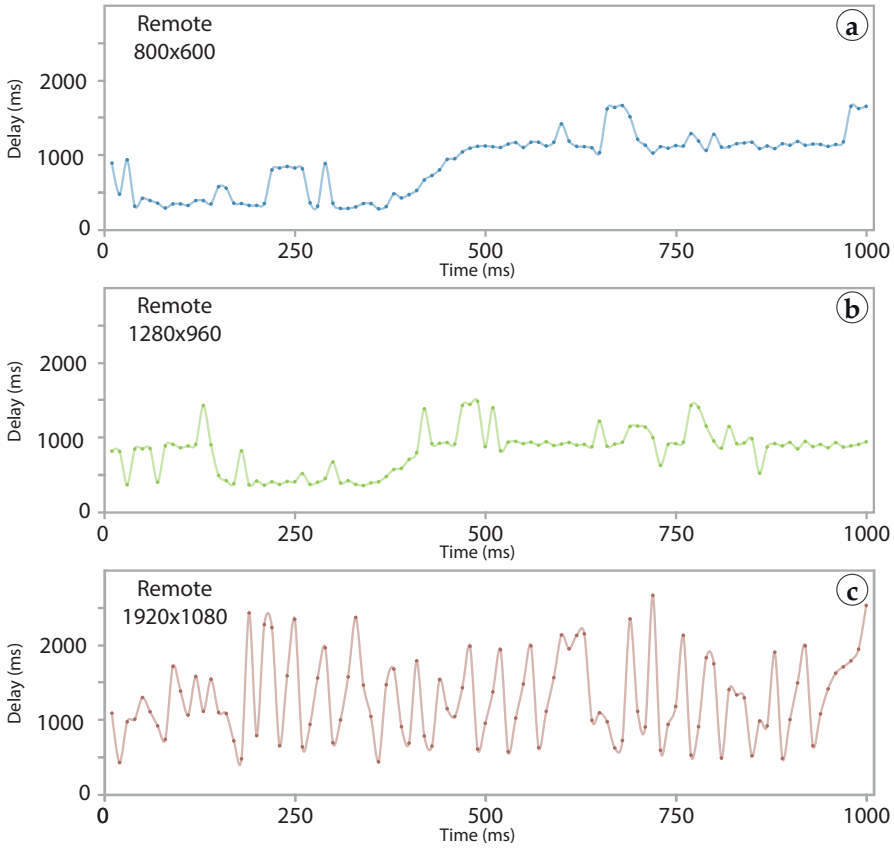


Figure 46: *CamRay* remote performance test result with 3 different resolutions: (a) 800 × 600, (b) 1280 × 960 and (c) 1920 × 1080.

The *Capture Module* and *Dispatch Module* were running on *frontal*1 in *WILD* and *master*1 in *WILDER*.

For a resolution of 800 × 600 (Figure 46 a), the mean delay is 902± 414ms, with a median of 1085ms. The delay went up occasionally by about 400ms for a short period, although half way through the process it went from about 380 to about 1100 and stabilized at that value. For a resolution of 1280 × 960 (Figure 46 b), the mean delay is 829±54ms with a median of 890ms. The delay went up occasionally by about 500ms for a short period, although, as before, half way through the process it went from about 400ms to about 900ms and stabilized at that value. For a resolution of 1920 × 1080 (Figure 46 c), the mean delay is 1295±567ms with a median of 1110ms. The delay was really unstable and kept going up and down, from about 600ms to 2250ms.

The delay for the first two resolutions is about 1 second after it stabilizes, although for 1280 × 960 the video sometimes has some artifacts, such as having fewer colors. For 1920 × 1080 the delay increases and decreases continuously, and the video sometimes has fewer colors.

The artifacts come from the fact that *WebRTC* adapts the resolution and compresses according to the available network bandwidth.

With these results, I choose to work with 800×600, as the delay is acceptable after stabilizing and there are no artifacts from the protocol adapting to the network load.

5.3.6 *Scaling CamRay*

As new cameras with higher resolution become available on the market the processing and network load will inevitably increase. To scale the current deployment of *CamRay* and cope with a higher video resolution, we can improve mainly 3 aspects: the hardware where each module executes, the network, and the video encoding algorithm.

Currently, the computer running the *Selection Module* in one room also runs the *Display Module* that receives data from the remote room. Because *CamRay*'s modules are independent, they can be deployed on dedicated computers. Thus, the first measure to mitigate a higher processing load would be to run each module on an independent computer with a high processing power.

Regarding the network, currently both *WILD* and *WILDER* have a dedicated network and are connected using a local network through a Virtual Local Network (VLAN). This infrastructure could be improved using a dedicated connection between the rooms. Although it is hard and costly to have dedicated networks between distant rooms across the globe, it is still possible if needed. There are protocols that aim at reserving network resources in remote communication, such as the Resource ReSerVation Protocol (RSVP), which allows network resources to be reserved to ensure an end-to-end quality of service [16].

Lastly, regarding the technology used for encoding, in the future better compression algorithm for sending media will make it possible to send higher resolution video with a lower bandwidth.

5.3.7 *Maintaining CamRay*

The most difficult challenge in maintaining *CamRay* is keeping all the technological components up to date. H.264, *WebRTC*, *OpenWebRTC*, *OpenCV* are all technologies whose implementations are constantly evolving. As new releases become available, APIs change and functions are sometimes removed. Even features that are meant to be transparent, often times require re-coding when they are updated, such as the Secure Socket Layer (SSL) protocol which is used when sending video across networks. It is challenging to keep *CamRay* up to date with the most recent releases of each software component, and ensuring that they work together seamlessly. Most of the time, the changes are confined within each module, as the definition of the protocols used for the communication between modules does not change as often as their actual implementation.

16 https://tools.ietf.org/rfc/rfc2205.txt

5.3.8 *Extending CamRay*

The flexible architecture of *CamRay* allows for new cameras and display surfaces to be added easily. To add a new camera, it suffices to send an RTP H.264 stream to the *Selection Module*. Apart from the *VICON* data, the *Selection Module* could be adapted to use other sources to select which camera should be active, such as a depth camera (e.g. Microsoft Kinect). It could also use image processing to analyze all feeds and select one based on their content, for instance when it detects a face. With the current implementation, this can be easily achieved as this module already uses *OpenCV* to receive and decompress video feeds.

To add a new display where video is presented, it suffices to add it in the array describing the tree structure so that the video feed is forwarded to it. The *Display Module* is based on web technologies, which makes it possible to add virtually any type of display as most devices with a display (e.g. tablets) now include a web browser.

This architecture makes it possible to support more than one user and locations. For multiple users, the motion tracking system can feed data of their position to the *Selection Module*, which can in turn select and forward their individual streams. For multiple locations, the *Selection Module* can forward streams to multiple sites, and the *Dispatch Module* in those locations can receive these multiple streams and forward them to the display individually. More than one *Display Module* can run in each screen to handle multiple video feeds.

I envision in the future situations where devices are added in a plug 'n play manner to prototype and test new interaction possibilities that target diverse collaborative practices.

Mobile devices for instance could be used as "on demand" cameras and displays. Collaborators could use their phone as a mobile camera that they can manipulate and show remote people physical objects when needed. This possibility has shown benefits for remote repair tasks where a person with broken hardware can get remote assistance from an expert [84].

Another interesting scenario could include telepresence robots that embody remote collaborators and move according to their position. We can use the robots' display to present the face of the remote collaborator, which is captured from the front by the camera array. These robots can also use a camera to capture video, which could be presented on the remote wall-sized displays. Yet another possibility could be to add drones to have a camera that flies around the user, capturing her face from any angle.

In short, *CamRay*'s flexible architecture and use of open technologies make it easy to add new cameras and display surfaces that allow for exploring how to support diverse collaboration practices.

5.4 SUMMARY AND CONTRIBUTIONS

CamRay is a telecommunications system for wall-sized displays. It uses an array of cameras to capture people as they move in front of a display, and presents this video feed on a remote display with the ability to move it across monitors controlled by different computers. It runs independently of the applications on the display, allowing *CamRay* to be used in a variety of applications.

CamRay can scale along several dimensions: the size of the display, the number of users at each site, and the number of sites. *CamRay* makes it easy to add more cameras as a tiled display grows in width, and the tree-shaped distribution scheme can be adapted to include more display surfaces. If the controlling computer becomes overloaded due to a larger number of cameras or higher-quality video, the load can be distributed over several computers, each one connected to a subset of the cameras. As more users are present in the room, they can be identified using the *VICON* system, and the *Selection Module* could select more than one stream to forward. Lastly, as more locations are added, the *Dispatch Module* in each site could simultaneously receive several *WebRTC* video streams and forward them to the *Display Module*s running on the display tiles. In this multi-user, multi-site environments, *CamRay* would send one video stream per user and per site, which is still much less than the total number of cameras per site. Also, the *WebRTC* protocol traverses almost any network, making it possible to connect to diverse sites with different network infrastructures, even over firewalls.

CamRay can capture and display video in flexible ways, making it a platform to study how to support remote collaboration across wall-sized displays by testing different combinations of video input and output during collaborative tasks. In the next chapter, I carry out this exploration.

6

WHERE TO DISPLAY VIDEO?

This chapter reports on a semi-structured observations where video is captured and displayed in various ways during two tasks, one that benefits from pointing gestures and the other from face-to-face conversations. I observe how different ways of displaying the remote collaborator's video incurs trade-offs in communication. The observations suggest that when video follows the remote participant's movements, pointing gestures are better understood, and when video follows the local participant's movement, face-to-face gestures are more frequent.

6.1 INTRODUCTION

CamRay lets us explore how to support the needs for remote collaboration across wall-sized displays. In particular, it lets us study how capturing and displaying video of remote collaborators in various ways supports different collaboration practices.

Chapter 2 presented the advantages of wall-sized displays for various individual and collaborative tasks, which come from the fact that these spaces promote physical navigation. Chapter 3 showed that wall-sized displays lead collaborators to point towards information and hold face-to-face conversations. To evaluate how to support these two practices, I conduct semi-structured observations where I manipulate how video is captured and displayed across two wall-size displays, while two collaborators work on a task at a distance. I analyze how each technique supports different needs for communication.

6.2 EXPLORING THE DESIGN SPACE

I start by exploring where to capture and display collaborators' video as they work across wall-sized displays.

For capturing video, a fixed or mobile camera can be used. A *fixed camera in front of users*, placed on the wall-sized display, can capture users' faces and upper bodies, as people normally stand close to the screen. But since users move in front of the display, we still need to consider how to follow them. A *fixed camera* on the *side* or *back* provides a view of users' bodies in context of the displayed information, from the side or back. A *mobile camera* can be implemented using a *telepresence robot*, covering the 2D space in front of the display, or a *flying drone*, covering the volume of the room.

For displaying video, there are also different possibilities: video can be *in front of the user on the wall-sized display* or *on the side on an independent screen*. Presenting video on the wall-sized display can still be achieved in multiple ways: it can *reflect the local and remote user*

interaction by moving according to their position, *or be static at a fixed location* on the display to provide contextual information. Also, video can be displayed on a *telepresence robot* or even possibly on a *flying drone* with a screen attached.

These input and output spaces can be combined, although some combinations make more sense than others. Table 1 shows these combinations.

					Capture			
					Fixed			Mobile
					Front	Side	Back	
Display	Fixed	Wall	Dynamic	Remote User	★			☆
				Local User	★			☆
			Static			☆	★	☆
		Side				★	☆	☆
	Mobile				☆			☆

Table 1: Combinations for video capture and display for wall-sized displays. Combinations that make sense and are explored in this chapter are indicated with a filled star (★); combinations that make sense but are not explored are indicated with an outlined star (☆).

Video of a user captured from the front shows a person's face and upper body. This contains detailed information, such as someone's gaze, facial expression or emotions. It makes sense thus to display this video in front of collaborators and not on the side, analogous to having two people stand in front of each other, rather than on the periphery. As both collaborators move, presenting video on the large display is a challenge. I propose for this study that video follows either the local user, making it always available, or the remote user, conveying his position. I call these combinations *Follow-Local* (Figure 47 top) and *Follow-Remote* (Figure 47 bottom) respectively. Another possibility would be to present this video on a mobile robot or a flying drone with a screen, so that it can be brought to the user as needed. Although displaying video on robots or drones might bring new opportunities for collaboration, I do not consider it in this study.

Video of a user captured from the side and from the back shows actions in the context of the displayed information, such as pointing at an object or moving to navigate data. This video contains fewer details than video of a person's face, thus there is not much interest in presenting it in front of collaborators, but on the periphery of their focus of attention, so that they can access this resource when necessary. I propose for this study that video captured from the side is presented on the side, which is similar to when people stand side-by-side to work on a problem together. I call this combination *Side-by-Side*. I also propose that video captured from the back is displayed statically on the screen, so that it can be used as a resource when collaborators need to interpret actions in the context of the displayed information. Due to the risk of occluding important content, and since the video

from the back is meant to provide context, displaying it at the bottom seems an appropriate solution. I call this combination *Back-Video*.

Video of a user captured from a mobile camera, either via telepresence robots or flying drones, would provide the flexibility of having close-ups as well as medium shots from any angle. It makes sense thus to present this video on any display: in front of users on the wall-sized display to show close-ups; as static video on the wall-sized display or on the side to show users in the context of the room; or even on a mobile robot or drone to bring video close to the observer when needed. Presenting video on a mobile device, either a robot or a drone, could be done in the same way as on the wall-sized display: following the remote user, thus conveying his position, or following the local user, thus making video available as the task requires. The variety of ways in which video can be captured by a mobile camera, makes it interesting to explore displaying it on any surface. In this work however, I do not explore displaying video on robots nor drones.

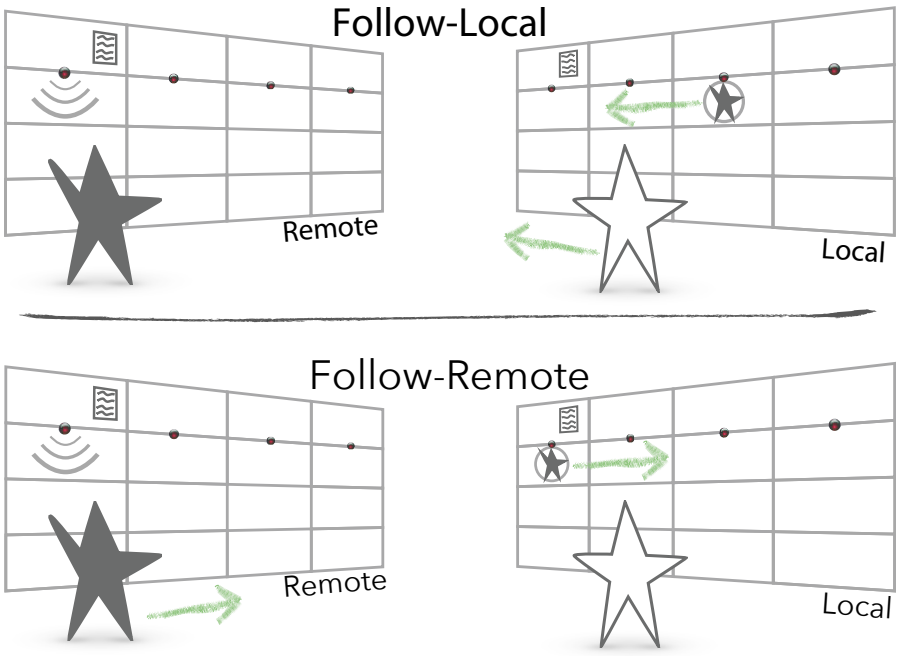


Figure 47: (top) *Follow-Local* and (bottom) *Follow-Remote* conditions. In *Follow-Local*, the remote person's video moves according to the local user position. In *Follow-Remote*, the remote person's video moves according to the remote user position.

6.3 OBSERVATIONAL STUDY

6.3.1 *Method*

The study is a semi-structured observation, where dyads perform a collaborative task at a distance using a mix of different CONDITIONS. The techniques include the combinations of capturing and displaying video in the room, and two control conditions, one where collabora-

tors have no video (only audio) and another where they are in the same room.

1. *Follow-Local*: the remote collaborator's video is captured on the remote display using *CamRay* and presented on the local display according to the local person's position;
2. *Follow-Remote*: the remote collaborator's video is captured on the remote display using *CamRay* and presented on the local display according to the remote person's position;
3. *Back-Video*: video is captured from the back of the remote room and displayed at the bottom of the local wall-sized display;
4. *Side-by-Side*: video is captured from the side of the remote room's display and displayed on a Microsoft Surface Hub display, positioned at a 90° angle on the side of the local display.
5. *Audio-Only*: there is no video but only an audio channel; and,
6. *Co-located*: both participants are in the same room.

In *Follow-Local*, the remote collaborator's video follows the local collaborator's position, thus each side sees the remote partner's video always in front of them. With this condition I hope to support face-to-face conversations, as suggested by the observations in Chapter 3 where collaborators used the front-facing video when talking about the content of objects.

In *Follow-Remote*, the remote collaborator's video follows the remote collaborator's position, thus each side sees the remote partner movements in their space. Here, pointing actions towards objects might be easier to interpret as they are displayed congruent with content, keeping their spatial relation. On the other hand, it may happen that one person is on the far left of his room while the other one in the far right, thus they don't see each other's video.

In *Back-Video*, the remote collaborator's video is captured from the back and displayed at the bottom of the screen. Observations in Chapter 3 suggest that pointing gestures might be better supported, as collaborators used the back-facing camera to obtain context when interpreting their partner's deictic gestures.

In *Side-by-Side*, the remote collaborator's video is captured and displayed on the side, without moving.

In *Co-located*, both collaborators are in the same room. Although this condition does not include *remote* collaboration, it might shed light to collaborative practices that need support when video-mediated.

In *Audio-Only*, there is only an audio connection, without video. This condition might show participant's struggle to communicate information at a distance when they cannot see each other.

I hypothesize that each of these conditions will facilitate different moments in communication. Although in principle *CamRay* can display the video freely around a tiled display, I choose to make the video switch discretely across monitors for *Follow-Local* and *Follow-Remote*, such that it is always right below a camera. This gives the

impression of direct eye gaze since when people look at the video, their gaze is right below the camera capturing their face, as discussed in Chapter 4.

6.3.2 *Participants*

Six pairs of people worked on two different tasks. All of them were researchers at our research team, 4 were PhD students, one was a post-doc and one an assistant professor. All performed the tasks in English.

6.3.3 *Hardware and Software*

The observation took place across *WILD* and *WILDER* using *CamRay* as well as a 84" Microsoft Surface Hub, located on the side in each room (right on *WILD*, left on *WILDER*). I used *Webstrates* [69] to display content during both tasks and synchronize it across displays.

6.3.4 *Procedure*

The six dyads collaborated using *Follow-Local* and *Follow-Remote*, four used *Co-located* (D3, D4, D5 & D6), three *Audio-Only* (D4, D5 & D6) and two *Side-by-Side* (D5 & D6). Each CONDITION lasted about about 10 minutes.

In all the CONDITIONS dyads could hear each other, and, except for *Audio-Only* and *Co-located*, a video feed displayed at bottom center of each wall-sized displayed showed the remote room captured by a camera at the back.

6.3.4.1 *Task Description*

I designed two tasks in order to observe different types of collaboration. The tasks purposely reflect dyads' common activities, both to maximize their involvement and to reflect real collaboration practices.

Task 1: Literature Classification: research articles from the dyads' collection of references for a future publication are laid out on the displays. Both participants hold iPads to enter text in four areas of the screen that correspond to four digital post-it notes. I use a Wizard-of-Oz technique to interact with the articles. An investigator sitting in one of the rooms can move the post-it notes and articles, as well as turn pages when requested to do so. This task requires participants to walk back and forth between articles, showing each other details as well as the relations among them.

Task 2: Research Discussion: the two collaborators discuss their own research. They must come to an agreement and write down their contributions along three dimensions using virtual post-it notes: theoretical, technological and empirical. For this task, collaborators also hold iPads where they can enter text in three areas, which correspond to the dimensions. The investigator playing the Wizard-of-Oz role can

create new post-its, move them across the display and change which post-it the participants can edit on their tablets.

Two dyads performed Task 1 (D1 & D2) whereas the other four (D3–6) performed Task 2.

6.3.5 *Data Collection and Analysis*

Sessions were video recorded in both rooms. After the dyads had completed their task, we held a debriefing interview where I asked them about how they used each technology, the problems they encountered, their preference and general impressions. These interviews were audio recorded.

I watched the videos from the sessions and listened to the interviews to perform thematic analysis [17]. I first identified interesting moments, such as referring to objects by pointing with the hand or pointing with the head, or looking at the video of their partner. Then I extracted themes based on those moments, which I present below. Lastly I watched the videos a second time to review the initial coding and to make sure there were no events omitted in the coding.

6.4 RESULTS AND DISCUSSION

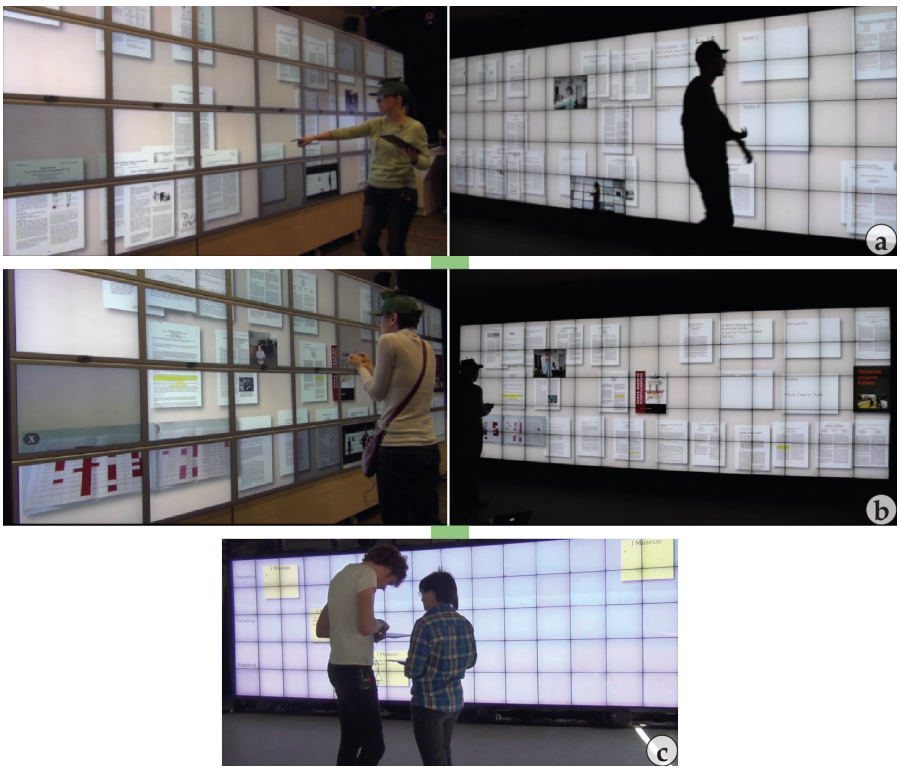


Figure 48: Observational Study with *CamRay*, dyads 1, 2 and 3. (a) D1: the left participant's pointing action can be seen in the right side video in the *Follow-Remote* condition. (b) D2: the left participant describes an article using hand gestures in the *Follow-Local* condition. (c) D3: both participants edit text in a post-it note using an tablet in the *Co-located* condition.

Each task was useful to understand how collaboration takes place across wall-sized displays using different techniques. As expected, in the literature classification task, participants went back and forth between articles while indicating to the remote collaborator which one they were currently talking about. In the research discussion task, participants worked on one post-it at a time while holding long discussions When presenting results, D1–6 refers to the dyads, while e.g. D3.1 and D1.2 refer to the individual participants.

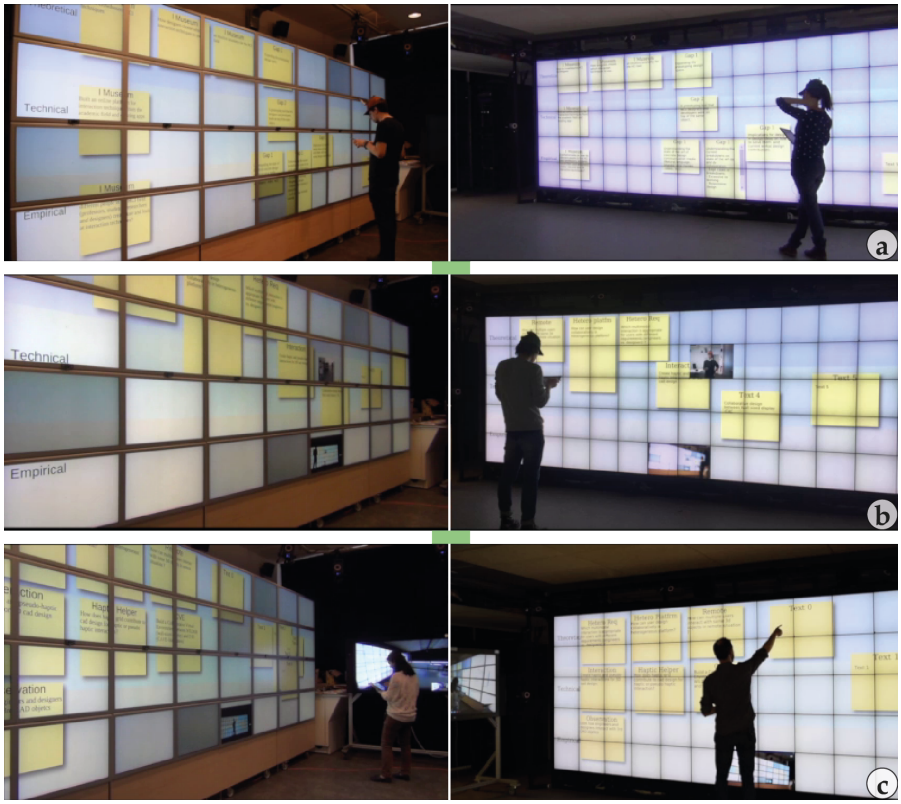


Figure 49: Observational Study with *CamRay*, dyads 4, 5 and 6. (a) D4: both participants discuss over audio the content to write on a post-it note in the *Audio-Only* condition. (b) D5: the right participant uses the back video (*Back-Video*) to determine his partner's position. (c) D6: the right participant points to a post-it while the left participant writes on a post-it note using her tablet in the *Side-by-Side* condition.

6.4.1 *Use of Deictic Gestures*

Participants used deictic gestures in all CONDITIONS.

In *Follow-Local*, pairs often misunderstood them, most notably D1 & D2, as the literature classification task inherently required more pointing actions. Participants typically mitigate this using two strategies: either they (1) synchronized their positions so that the video is in the correct position to interpret the pointing action, or (2) use more verbal explanations to provide context. Although they could have used the back video to understand these actions, they rarely did so. D2.1 explicitly mentioned that it takes an extra effort to look at this video.

In *Follow-Remote*, deictic gestures were better understood, for instance when D1.1 says *"it's next to the cluster here"*, D1.2 replies *"yeah"* to confirm that he understood where "here" is (Figure 48 a). D2 and D3 understood more instructions in this condition than in the others, while D4, D5 & D6 did not use many deictic gestures during the task.

Side-by-Side allowed participants to produce and interpret deictic instructions to a certain extent (Figure 49 c). D5.2 for instance mentioned that he can determine not only the position of the remote person from the side video, but also where he is pointing.

Audio-Only generated the most confusion, so dyads turned to lengthy verbal descriptions of objects when referring to them. D4.2 admits that he verbalizes more which object he is talking about, as there is no video to convey this information (Figure 49 a).

When *Co-located*, there were no problems interpreting deicitc gestures, as expected.

6.4.2 *Discussions*

Holding face-to-face discussions happened mostly in the research discussion task (Task 2). Still, for both tasks and in all conditions, participants performed hand gestures to augment their speech while discussing.

Follow-Local supported conversations even when participants were standing on opposite sides of the rooms. This means that participants could talk and see each other's faces without having to change their positions. This allowed them to talk about two objects that were related but located far away (Figure 48 b).

In *Follow-Remote* participants followed each other's video to see their faces while they held discussions. This behavior was consistent across participants, and in some cases so strong that even when one participant (D1.2) was thinking out loud—talking to himself—and walked back and fort, the remote person followed the video. Some participants used the video to figure out the position of the remote person and what was the current focus of their attention: D5.2 mentions in the interview, *"I can see [D5.1] was here, so I understand he's here and we are talking about this memo"*.

In *Side-by-Side* participants got closer to the side display to talk to each other, while in *Audio-Only* participants produced fewer hand gestures, probably because, as D6.1 mentions in the interview, they are aware of the limitations of the channel. When *Co-located* participants looked at each other and stood next to each other while talking (Figure 48 c).

6.4.3 *Content Occlusion*

Follow-Remote was the condition in which content occlusion caused the biggest problems, as participants were not in control of the position of the video. To mitigate this, participants usually instructed their partners to move, for instance D3.1 asked her partner *"Can you*

go away? Because you're exactly on my post-it". D4.2, asked the Wizard-of-Oz to move a post-it as the remote video was on top of it to avoid bothering her partner.

In *Follow-Local*, occlusion occurred often (D1–4), but since the video is controlled by the local person, participants mitigate this problem by moving left or right. They do not need to ask their partner to move. Content occlusion was not a problem in the other conditions, as there was no video on the display.

6.4.4 *Use of the Back Video*

The back video was mostly ignored in all conditions, it's rare to see participants glancing at the back video or to talk about it. In the interviews, all the participants mentioned that, although they see the benefit of the back video (Figure 49 b), they do not really use it. For example, D1.1 mentioned that she liked the back video *"but in the second condition [Follow-Remote] it was fine"*, meaning that she did not need it as the location of the remote partner is already conveyed by the front video. When asked if she used the back video in *Follow-Local*, D3.1 replies that she only used it to count the number of screens on the remote side to locate some content and compare it to her display, although she was aware of this video's benefits: *"Not really, I saw it when he was asking how many screens there are [...] I noticed it, and I noticed what it was doing"*.

6.4.5 *Awareness of Partner's Activities*

Some dyads (D1–3) mentioned that they used the video not only to show gestures, but also to check what the other person was doing. D3 mentions: *"The first one [Follow-Local], it was easier in the sense that I could see him writing, so I didn't have to ask: are you writing this one?"*.

In both *Follow-Local* and *Follow-Remote*, it seems that the video window was seen as a proxy for the remote collaborator. When a person was out of range from the field of view of the cameras, their partner would continue to talk to the video feed showing the back of the room. This happened for instance when D2.2 sat down on the floor because of fatigue, or when D1.2 knelt down to read content at the bottom of the display.

6.4.6 *Self Video Feedback*

I was intrigued to find out if people were bothered because they could not see their own faces, as it is the case in most video conferencing systems. None of the participants mentioned that this was a problem. For instance, when asking D2.2 explicitly about this, she explains: *"I didn't ever though about, oh, the fact that there wasn't, you know, the typical Skype window of yourself that you can see, so you can kind of fix your hair, make sure that your teeth are clean and all that kind of stuff, [...] I just didn't even kind of realize, oh, there is no video of myself to check myself and*

I think it's because this just feels much more immersive than a teleconference situation, where you feel uncomfortable [..] here, I feel like well, I don't need to frame, like, I'm just trusting the technology is capturing me properly, where I am, and that the window is not, you know, getting my chest instead of my head or whatever it is".

6.4.7 *Preference*

Participants were asked to rank the remote communication media they used during the interview.

Follow-Local was preferred by 6 participants (D3, D4, D5.1 and D6.2), and also liked by another two (D2) even though it was not the preferred one. Both D4.1 and D4.2 prefer this condition, D4.2 mentions: *"because also, we were tackling one item at a time, so when we were talking about one sticky note, we were talking about that sticky note, so [...] the position in the wall wasn't so important"*. D3 prefer this condition although D3.1 mentioned later in the interview that *"In the first condition [Follow-Remote] it was much more intuitive to know... to look at where he was, and where he was going, and what he was talking about"*. D6.2 mentions at one point of the interview that he prefers *Follow-Remote* although he is not in control of the video. He also admits that *Follow-Local* is better, and changes his mind later on and ranks *Follow-Local* as his preferred condition. It is interesting to note that all the participants that liked this condition worked on *Task* 2.

Follow-Remote was preferred by 5 participants (D1, D2 and D6.1). D6.1 mentions that she prefers it over *Follow-Local* because it is overwhelming to see the other person's face all the time, even if she feels more in control of the video position. She also mentions that this condition gives her a better feeling of where the other person is standing and where he's looking at, and that it can even be used for pointing. It is interesting to note that this condition was preferred by the two pairs that did *Task* 1.

Side-by-Side was preferred by 1 person (D5.2), because he likes seeing the position of the remote partner in the remote space.

All participants mentioned that *Audio-Only* was the least useful technique for collaborating. For instance, immediately after switching, D4.2 mentions *"I don't like this condition, can I get back?"*. Participants consulted each other often: *"Are you gone, [D4.1]?"* after D4.2 had asked a question, but the other side was silently thinking about a reply. D5.1 did not like this condition at all, remarking that he felt lonely. D6.2 mentioned that although it is the least preferred, the condition did not feel weird as it was the last one and he was already confident of what was happening in the remote side.

6.4.8 *Other Technology-Related Comments*

In *Follow-Local*, D3.1 mentioned that the camera switch caused by the motion of her partner is bothersome. This effect is more noticeable in this condition because the video is still. There is a dissonance be-

tween the fact that the remote person is moving, but their video is still. One artifact of *CamRay* that became visible was that when a participant stood in front of the edge where two monitors meet, the video controlled kept jumping back and forth between two monitors. This happened for every participant, and it was explicitly mentioned by D2.

6.4.9 *Task-Related Comments*

Participants were engaged during the task, mentioning in the interviews that they liked the wall-sized display and that they found it useful. Having a real-world task that participants were working on during the week of the observations increased the level of engagement. The first two dyads mentioned that the literature sorting task would have been very difficult to do on a desktop. D1.1 mentioned that, contrary to Skype, this setup was *"so much better"*, and D1.2 that *"something that I like from the wall is that you can see all the content at once"*. D2.2 thought that *"it feels very different than a Skype session"*, and that *"I can't image doing this [by Skype] I mean, [...] we had a lot of discussion just about the framing of the paper [...], and that can be done by Skype. But [...] for me to kind of visualize the papers that you're thinking about as being in the related work, [...] I don't know how you'd do that by Skype. To me, this was much more immersive"*. Still, one drawback of the wall-sized display is fatigue. Two participants felt tired from standing all the time, one of them sat down during the task and the other asked permission to do so.

6.5 SUMMARY AND CONTRIBUTIONS

For the Task 1, *Follow-Remote* seems more adequate as there is a need to know where the remote person is standing to interpret their deictic gestures. This technique is an implementation of Buxton's *reference space* [21].

For the Task 2, *Follow-Local* enabled face-to-face discussions to take place even when collaborators stood in different sides of their corresponding displays.

It seems that these two techniques support different types of communication, but also change the way in which collaborators communicate. For example, when referring to on-screen objects, in *Follow-Remote* participants simply pointed and said *"here"*, whereas in *Follow-Local* they had to be more verbose, or walk towards the object and repeat the gesture. Also, checking for the remote person's understanding was easier in *Follow-Local*.

Video on the side (*Side-by-Side*) had the benefit of not occluding on-screen content, but at the cost of turning one's head and loosing view of content while walking towards the display during a discussion.

Based on this observational study, I made the following changes to the setup used in the experiments described in the next chapter

- removing the back video, as it was not used even in the conditions where video did not convey the speaker's position;
- setting a threshold before switching the camera that captures a person, to prevent the video window from switching from left to right back and forth when a person is standing at the edge between two monitors; and,
- removing the self video feedback as there is no apparent need for it.

This study suggests that having the video follow the remote collaborator (*Follow-Remote*) is better for tasks where collaborators manipulate shared digital content, and refer to it using pointing gestures. This setup provides the benefits of capturing video from the back, but with the benefit of relying solely on *CamRay*'s array of cameras. This study also suggests that having video follow the local collaborator (*Follow-Local*) is better for tasks that rely on conversation about the content of shared digital objects.

In the next two chapters, I perform experiments to verify these observations in a controlled environment, and to further explore the benefits and drawbacks of these two video behaviors.

7 SUPPORTING POINTING GESTURES

This chapter performs an experiment to validate the observations from Chapter 6 that making video congruent wit shared content benefits the correct understanding of remote pointing gestures. Pairs of participants perform a data manipulation task that heavily relies on pointing gestures. The results show that participants are able to manipulate data more efficiently, take less time, make fewer errors and use fewer words when video followed the remote collaborator.

7.1 INTRODUCTION

In this chapter I investigate the support for remote pointing gestures using *CamRay*'s two video behaviors that leverage the remote and local collaborators' positions. According to Fussel et al. [42], people devise novel strategies for pointing when technologies make it difficult to do so through natural hand gestures. I am interested in investigating the strategies people use for pointing when video is displayed according to the different strategies identified in the previous chapter, as some might not support pointing directly.

I conducted an experiment using a data-manipulation task that heavily relies on pointing, with the goal of understanding in detail how this systems supports the production and interpretation of pointing gestures. This chapter is based on work published at CHI 2017 [5] in collaboration with Wendy Mackay.

7.2 BACKGROUND

7.2.1 *Previous Work on Remote Pointing Gestures*

Previous research supports my observation of the need for pointing during collaboration using wall-sized displays. Inkpen et al. [52] observe large numbers of pointing gestures on large displays during a collaborative task, both when users stand close to and far from the display. Jackobsen & Hornbæk [62] observed that visual attention is eased by the fact that people stand close to the wall and can see each other's gestures. Liu et al. [76] observed that pairs are aware of their partner's positions and actions, such as pointing, and this prevents conflicts.

Remote pointing has been achieved before by integrating video and content. I have already discussed Buxton's [21] integration of *person* and *task* spaces into the *reference* space, where the remote person can use body language to reference the work—such as pointing gestures.

The first systems that implemented this integration are VideoDraw [106], VideoWhiteboard [107] and Clearboard [61]. They overlay an image of the remote collaborator with a shared drawing space, supporting pointing at parts of the drawings. Hydras [92] support pointing at remote collaborators in a multi-way remote conversation by keeping spatial relations consistent across sites.

This integration has also been explored in tabletops and whiteboards. Three's Company [104] integrated shadows of users arms on top of shared content, and ImmerseBoard [55] deformed the user's arm to place it on top of content during remote collaboration in a large vertical display. Authors have explored this relationship in collaborative virtual environments (CVE) [116] and by using egocentric view points [1]. In wall-sized displays however, the relation between video and content to support remote pointing gestures has not yet been explored.

Previous work have explored the combination of the remote person's representation and shared content to support pointing. I draw inspiration from how these systems achieve this combination to study remote pointing gestures for collaboration with shared data across wall-sized displays.

7.2.2 *The Cost of Mediating Communication with Technology*

In Chapter 2 I quickly reviewed Clark's [28] Common Ground theory. In short, according to Clark & Marshall [28, 29], communication is characterized by a series of messages between parties which, once understood, become part of their common ground: the mutual knowledge, beliefs and assumptions shared by partners in communication [28]. Common ground is updated through grounding, the collective process by which participants try to reach mutual belief that what has been said has been understood [26].

Technology-mediated communication has a cost, which highly depends on how the system helps people ground their utterances during communication. Clark defines and characterizes several costs for technology-mediated communication [26]. *Formulation costs* relate to the effort it takes to formulate and reformulate utterances. *Production costs* are associated to the effort of the act of producing an utterance. *Reception costs* are the cost of how information is received, for instance listening is generally easier than reading. *Understanding costs* relate to how hard it is to understand certain words, constructions and concepts. *Start-up costs* come from the difficulty to starting a new discourse. *Delay costs* are associated with the cost of delaying utterances to plan, revise it and execute it more carefully. *Asynchrony costs* come from media with delays in information transmission. *Speaker change costs* arise from the cost of switching speakers. *Display costs* relate to the limited ability to gaze, point to objects or nod at people in remote situations. *Fault costs* are associated with producing mistakes. And finally, *repair costs* come from repairing utterances.

I believe that telecommunication systems must take these costs into account and attempt to reduce them in order to support effective video-mediated communication.

7.3 STUDYING REMOTE POINTING GESTURES ACROSS WALL-SIZED DISPLAYS

I conducted an experiment to assess the effects of video position on communication and the trade-offs it incurs. The goal is to compare the two video behaviors for displaying a remote collaborator's video with *CamRay*, and a control condition:

- *Follow-Remote*: the video windows appears on the wall at the same position as the remote collaborator;
- *Follow-Local*: the video windows appears on the wall in front of the local user; and
- *Side-by-Side* (control condition): the fixed video window appears on a separate screen, perpendicular to the wall.

In Chapter 6 I observed how deictic gestures seemed to be better supported when the video was placed in the context of the shared content. Therefore I formulate the following hypotheses:

- *H1: Follow-Remote leads to the production of more deictic instructions than Follow-Local and Side-by-Side;*
- *H2: Follow-Remote is more efficient for data manipulation tasks than Follow-Local and Side-by-Side; and,*
- *H3: Follow-Remote is preferred for manipulation tasks when giving and receiving instructions.*

7.3.1 *Method*

The $[3 \times 2 \times 2]$ within-participant design has two primary factors and a secondary factor:

- VIDEO has three behaviors: *Follow-Local*, *Follow-Remote* & *Side-by-Side*;
- LAYOUT has to two difficulty levels: *Medium* & *Hard*; and
- ROLE (secondary) corresponds to the asymmetric roles in the collaborative task: *Instructor* & *Performer*.

LAYOUT controls the difficulty of the task, while ROLE accounts for the asymetry of the task, as described below. The $3 \times 2 \times 2$ conditions are counterbalanced across participants using Balanced Latin Squares. The order of the three VIDEO conditions is counterbalanced across pairs; for each video condition participants switch between the two ROLES. Finally, the resulting block alternates between the two LAYOUTS.

7.3.2 *Participants*

12 pairs of participants took part in the experiment, aged between 23 and 40, all with normal or corrected-to-normal vision, none color blind. Pairs were formed as participants were recruited, leading to 9 male-male, 2 female-male and 1 female-female couples. 1 participant used video conferencing systems on a daily basis, 8 on a weekly basis, 6 on a monthly basis, 5 on a yearly basis and 4 almost never.

7.3.3 *Hardware and Software*

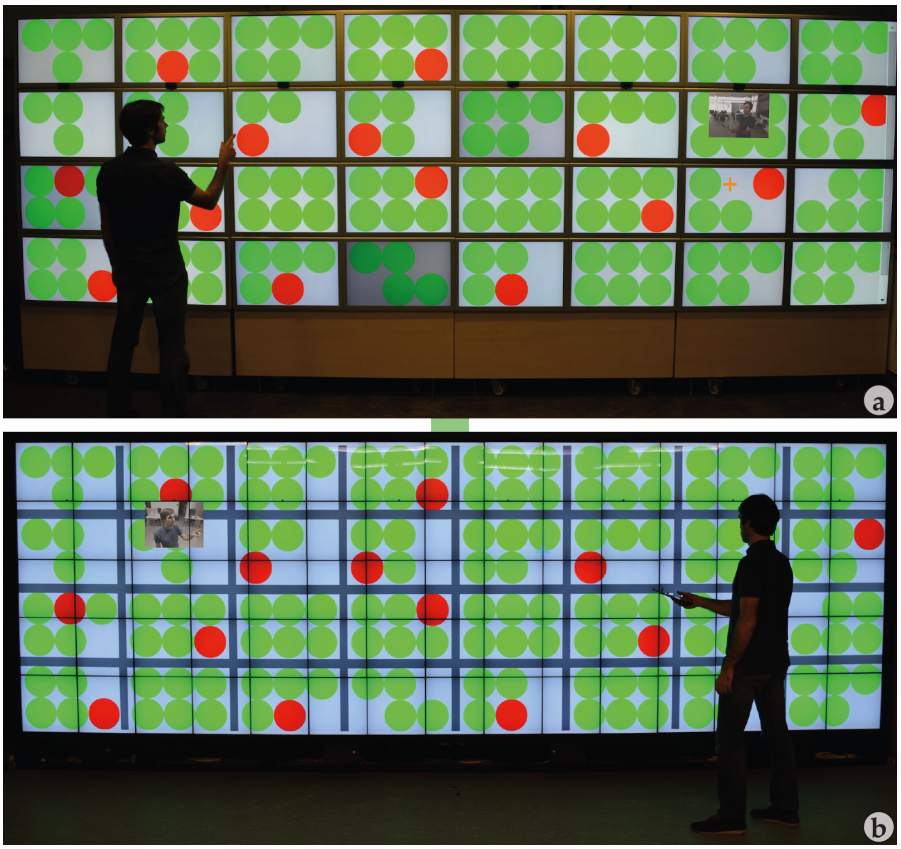


Figure 50: Experimental setup. (a) A participant in *WILD* pointing to a disc. (b) A participant in *WILDER* manipulating a disc.

The setup of the experiment is composed of the two wall-sized displays, *WILD* and *WILDER* (see Appendix A for details on these rooms). *Follow-Local* and *Follow-Remote* conditions are implemented with *CamRay* (Figure 50). The video windows of the remote users move horizontally at a fixed height of 1.75m (center of the window) at both sites, and are horizontally mirrored to preserve spatial references—a gesture performed by the right hand is seen remotely on the right-hand side of the video. In *Side-by-Side*, video is displayed on an LCD screen on the left side of the room, at approximately the same height as the window in the other conditions. In all three VIDEO conditions, the video windows have the same size (34.7cm $\times$ 26cm). Participants wear wireless microphones to capture audio, which is sent across

rooms through a *Google Hangout* call. Audio is played in each room using a high-quality surround sound system.

Although *WILD* and *WILDER* have different sizes and resolutions, the content is scaled so that it spans the entire display. *Webstrates* [69] is used to create and synchronize content. Participants interact with each wall-sized display with a cursor controlled by a handheld pointer through raycasting. The pointer is mounted on a smartphone that displays a button for picking and dropping. The orientation and position of the pointers and of the participants' heads are tracked using the *VICON* tracking system in each room.

7.3.4 *Procedure*

7.3.4.1 *Task Description*

As I am interested in remote pointing gestures using *CamRay*, the experiment task needs to operationalize the production and interpretation of such gestures.

I use a version of Liu et al.'s [75] task, which consists of classifying discs into containers based on their label (Figure 50). In one condition of their experiment, one participant had to tell the other which disc to move into which container. They naturally used deictic instructions, such as *"take this one and put it here"*. I adapt this abstract data manipulation task to a remote setting: at one site, the *Instructor* sees the labels and gives instructions, while at the other site, the *Performer* manipulates the discs, which do not display their label. This forces each dyad to produce and interpret deictic instructions.

Each wall-sized display is divided into 32 (8 rows $\times$ 4 columns) virtual containers holding up to 6 discs each (Figure 50). Discs belong to one of 8 classes, represented by the letters *C, D, H, N, K, R, X, Z*. When more than two discs of the same class are in the same container, they are properly classified and turn green. Misclassified discs are red. On the *Instructor* side, the disc classes are displayed in a small font (2mm $\times$ 2.5mm), forcing the *Instructor* to move to read the labels.

Layout: when the task begins, the layout features 160 discs, five per container. 12 discs are misclassified, distributed randomly across containers. The goal is to classify all the red discs by picking, moving and dropping them into a correct container. Each participant has a ROLE assigned: the *Instructor* sees the disc labels but cannot interact with them; the *Performer* sees green and red discs without labels but can manipulate them with a pointing device. The *Instructor* must therefore guide the *Performer* to classify the discs.

LAYOUT corresponds to task difficulty, achieved by varying the euclidean distance between a red disc and its closest solution[1]. This distance is between 1.5 and 2.6 for *Medium* layouts, and between 2.7 and 3.5 for *Hard* layouts. The further away a solution is from a disc, the more navigation is required, making the task harder. I generate random LAYOUTS for both *Medium* and *Hard* and pick one at random

1 The unit is the size of a container and the distance between two adjacent containers is 1.

when starting a new session. A trial corresponds to the correct classification of a disc, starting when the last disc of the previous trial is dropped, and ends when the disc is correctly classified (which may take several pick and drops).

Participants were welcomed and given paper instructions on how to perform the task. They were instructed to solve the task as quickly and accurately as possible. Participants filled a total of 5 questionnaires: one for collecting demographic data on arrival, one after each VIDEO condition, and one at the end of the experiment. Before each new VIDEO condition, participants performed 4 training trials (2 ROLE $\times$ 2 LAYOUT). Lastly, they could take a break after each LAYOUT and were encouraged to do so at the end of a VIDEO condition block. Sessions lasted about 70 minutes including the time to fill out the questionnaires.

7.3.5 *Data Collection*

Pick and drop event were logged with the time, position of the disc on the screen and number of discs left to classify. The *VICON* tracking system in each room recorded kinematic data of (a) user position, (b) user head direction and (c) cursor movement. Sessions were video recorded.

Pairs assessed their understanding of each other's actions and use of video in the questionnaire at the end of each VIDEO condition. The final questionnaire assessed the strategies and participant's preference when acting as *Instructor* and *Performer*. Questions were based on 5-point Likert scales and open-ended comments.

7.3.6 *Data Analysis*

I analyze three different measures: task performance, movement data from the kinematic logs, and conversations.

7.3.6.1 *Task Performance*

I chose Task Completion Time (*TCT*) as a measure of performance. The number of pick-and-drops for classifying one disc is a less useful indicator of performance than time, since all layouts were successfully solved with low error rates. *TCT* is the time required to correctly classify a disc. Since this may require several attempts, *TCT* starts when the *Performer* drops the previous disc and ends on the first drop in the correct container. I observed that some dyads picked one disc immediately and waited for an instruction, whereas others waited for an instruction and then picked a disc. To ensure a fair comparison and account for the time taken to find a container and produce the instruction in all trials, I include the time elapsed from the previous drop until a disc is picked in *TCT*.

7.3.6.2 *Kinematic Analysis*

To account for the slightly different sizes of the two wall-sized display, I normalize user position, cursor position and head direction between -1 and 1. After normalization, two users standing at the same relative position, e.g., the center of each room, have the same value, e.g., 0 on the X axis.

7.3.6.3 *Video and Speech Analysis*

Using the sessions' video recordings, I tagged each pick and drop and coded (I) the *Instructor strategy* to indicate containers/discs; (II) the *Performer error* when performing instructions; and, for both roles, (III) the *word count*, including the amount of deictic instructions.

I. Instructor strategy to indicate containers or discs was tagged using the following coding scheme:

- *Pointing*: using the finger to point, no verbal instruction;
- *Pure Deictic*: using only deictic instructions (*"..goes there"*);
- *Relative to Own Position*: deictic instruction relative to the *Instructor*'s position (*"here, one up"*);
- *Relative to Video*: deictic instruction relative to the *Instructor*'s video (*"where I am, second row"*);
- *Relative to Disc*: deictic instruction relative to where the *Performer* is moving the disc (*"there, one up"*);
- *Relative to Container*: deictic instruction relative to where the disc is picked (*"two to the right, one down"*);
- *Absolute*: relative to the display grid (*"column 3, row 4"*); and,
- *Based on Previous Pick/Drop*: using the location where the previous disc was picked or dropped (*"put it in the same place as the last one"*).

II. Performer error when performing instructions used the following coding scheme:

- *Understanding Error*: error when interpreting an instruction;
- *Instruction Error*: the *Instructor* provides a wrong instruction (the container is not of the same class as the disc); and,
- *Interaction Error*: the *Performer* accidentally drops a disc while moving it.

III. Word count serves as a measure of the efficiency when producing and understanding instructions. I used a coding scheme based on Gergle et al. [43]. I only coded utterances relevant to instructions, i.e. references to a specific disc and position. I counted words related to acknowledgment of behavior only when discs were not already dropped and changed their color to green; once this happened, words

were considered redundant for the classification and ignored. I ignored context information not relevant to an instruction (such as discussing the task itself) and back channel responses such as *"hmmm"* or *"so"*. Hauber et al. [51] also used this approach for counting words. Politeness forms were not coded, e.g., *"could you please"*. Finally, repeated terms were counted once, since I identified that many participants repeated utterances, e.g., *"that one, yes, yes, yes"*).

7.4 RESULTS

A total of 4330 pick and drop events were registered (excluding practice trials), aggregated into 1728 disc classifications (12 discs × 2 ROLE × 2 LAYOUT × 3 VIDEO × 24 PARTICIPANT).

For all measures, I perform a full factorial analysis[2] using the model VIDEO × LAYOUT × Rand(PARTICIPANT) using REsidual Maximum Likelihood (REML), unless otherwise specified.

7.4.1 *Task Performance*

TCT was tested for normality using a Shapiro-Wilk *W* test and that showed that it was not normally distributed. Kolmogorov's *D* for testing goodness-of-fit with a lognormal distribution yielded a non-significant result. Therefore, I ran the analyses using the log-transform of *TCT*, as recommended by Robertson & Kaptein [88] (p. 316). All analyses using the non-transformed time data yield the same effects with very similar p values.

The analysis of *TCT* yields a significant main effect of VIDEO ($F_{2,1699} = 7.69$, $p = 0.0005$), and of LAYOUT ($F_{1,1714} = 22.41$, $p < 0.0001$), but no VIDEO × LAYOUT interaction effect ($F_{2,1713} = 0.50$, $p = 0.61$) (Figure 51).

A post-hoc analysis[3] reveals that in *Follow-Remote* (13.23±6.88s), participants classified discs significantly faster than in *Follow-Local* (14.63±7.15s, p=0.0004) and *Side-by-Side* (14.41±7.47s, p=0.0173). There was no difference between *Follow-Local* and *Side-by-Side*. Data shows an improvement of *Follow-Remote* over *Side-by-Side* of 8.2% (1.19s); and over *Follow-Local* of 9.6% (1.41s).

2 All analysis are performed using SAS JMP

3 All post-hoc analysis are performed using a Tukey-Kramer "Honestly Significant Difference" (HSD) test

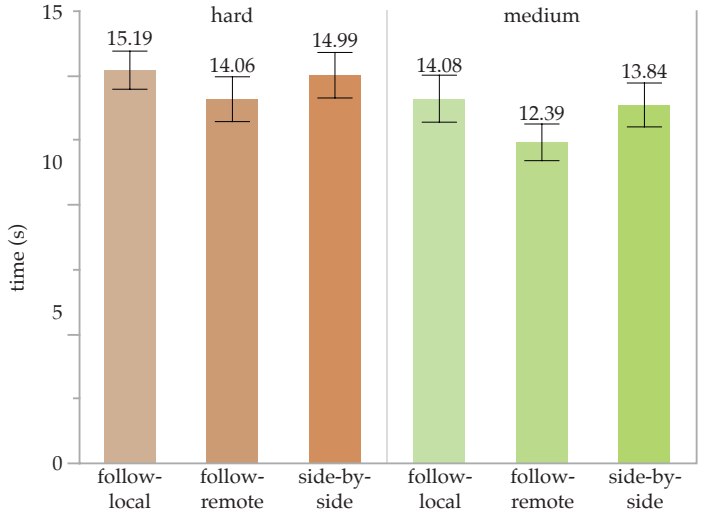


Figure 51: Time (*TCT*) in seconds for each VIDEO × LAYOUT condition. Bars show 95% confidence intervals.

7.4.2 *Kinematic Analysis*

In *Follow-Remote*, I observed that after picking a disc, the *Performer* would try to predict the target container by looking at the *Instructor*'s cursor and head direction. Some *Performer*s where even able to interpret target containers with minimal instructions, often following the *Instructor* and dropping the disc into the container that the *Instructor* was looking at.

I therefore compute two measures to further analyze this observation. *Cursor-Position Difference*: the horizontal distance between the *Performer*'s cursor and the *Instructor*'s position (Figure 52 a-c); and, *Cursor-Gaze Difference*: the horizontal distance between the *Performer*'s cursor and the estimated point the *Instructor* is looking at, based on the direction of the head (Figure 52 d-f). To get a single value per trial, I average these measures for all kinematic data points within the trial.

7.4.2.1 *Cursor-Position Difference*

The analysis of *Cursor-Position Difference* yields a significant main effect of VIDEO ($F_{2,1697} = 64.09$, $p < 0.0001$), and of LAYOUT ($F_{1,1711} = 32.42$, $p < 0.0001$), but no VIDEO × LAYOUT interaction effect ($F_{2,1711} = 0.023$, $p = 0.98$) (Figure 52 g). A post-hoc analysis shows that *Follow-Remote* (0.151 ± 0.116) has a significantly smaller *Cursor-Position Difference* than *Follow-Local* (0.228 ± 0.154) and *Side-by-Side* (0.227 ± 0.155).

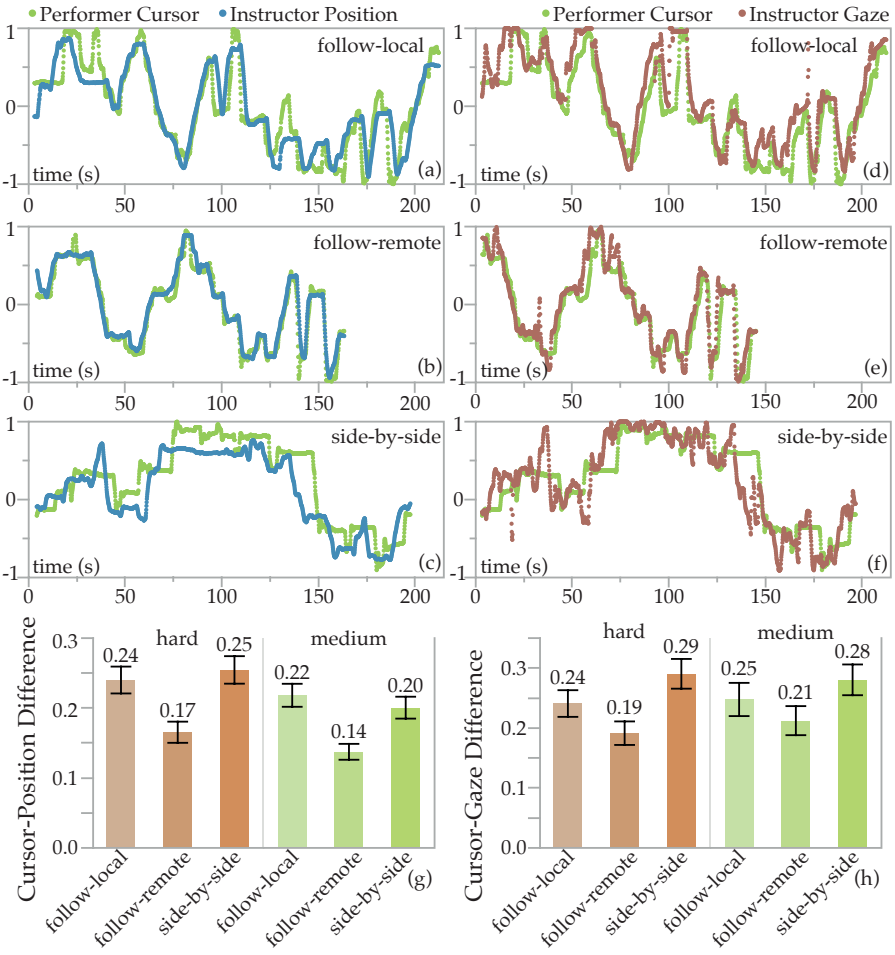


Figure 52: Kinematic data for dyad 9. The paths show a bird's eye view of the normalized horizontal positions of the participant, cursor and gaze over time. Left: *Performer* cursor and *Instructor* position for *Follow-Local* (a), *Follow-Remote* (b) and *Side-by-Side* (c). Right: *Performer* cursor and *Instructor* gaze for *Follow-Local* (d), *Follow-Remote* (e) and *Side-by-Side* (f). Histograms show the cursor-position difference (g) and cursor-gaze difference (h) for each VIDEO condition. Bars show 95% confidence intervals.

7.4.2.2 *Cursor-Gaze Difference*

The analysis of *Cursor-Gaze Difference* also yields a significant main effect of VIDEO ($F_{2,1697} = 30.64$, $p < 0.0001$), but not of LAYOUT ($F_{1,1704} = 2.28$, $p = 0.13$), nor of the VIDEO $\times$ LAYOUT interaction ($F_{2,1704} = 0.9021$, $p = 0.41$) (Figure 52 h). A post-hoc analysis shows that all VIDEO conditions are significantly different from each other. *Follow-Remote* has the smallest *Cursor-Gaze Difference* (0.201±0.190), followed by *Follow-Local* (0.244±0.216) and *Side-by-Side* (0.284±0.217).

7.4.3 *Video and Speech Analysis*

I analyze the strategies used by the *Instructor* and the errors produced by the *Performer* when picking and dropping discs. For this analysis, I use the tagged data for each pick and drop, not the aggregated data for correctly classified discs. While tagging video data, two new cat-

	Follow-Local	*Follow-Remote*	*Side-by-Side*
pure deictic	8	92	43
relative to video	3	318	0
relative to own position	0	0	7
relative to disc	116	61	148
relative to container	221	63	196
established pick order	284	302	257
based on previous drop	58	27	19
based on previous pick	10	5	12
absolute	447	254	435
arbitrary by *Performer*	80	83	111
none	248	241	247
total	1475	1446	1475
	4396		

Table 2: *Instructor strategy* for indicating objects.

egories emerged: *arbitrary* (no instruction), when the *Performer* picks any disc; *established pick order*, when the *Performer* picks in an order defined at the beginning of the session.

4396 events were tagged (Table 2), evenly distributed among the three VIDEO conditions. Note that events that have no strategy come from the *Performer* correcting errors (most often due to a failed interaction, such as releasing the disc too soon while moving it), which required no instruction: the *Performer* would re-pick the disc and drop it in the planned destination.

7.4.3.1 *Instructor Strategy*

I investigate the role of video on the use of deictic instructions by the *Instructor* (*Instructor strategy* or *IS*). I consider as deictic instructions the following categories: *Pure Deictic*, *Relative to Video* and *Relative to Own Position*, as the *Performer* needs contextual information that he cannot have on its own to understand the utterance. *Relative to Disc* and *Relative to Container* are also deictic in nature, but the *Instructor* has all the contextual information available (the disc he is manipulating and the container where he picked up the disc) without the need to consult the *Performer*. For the last two, I always observed that the *Instructor* used a deictic pronoun to make a reference relative to the position of the video or to herself. In *Follow-Remote* participants used 410 deictic instructions (318 *Relative to Video*; 92 *Pure Deictic*), 50 in *Side-by-Side* (7 *Relative to Own Position*; 43 *Pure Deictic*) and 11 in *Follow-Local* (3 *Relative to Video*, 8 *Pure Deictic*).

As expected, participants were able to use more deictic instructions in *Follow-Remote* (28% of total) than *Side-by-Side* (3.4%) and *Follow-Local* (only 0.7%). If we take a closer look at the strategies for disc drop events only, the use of deictic instructions in *Follow-Remote* goes up to 45% (265 relative to video, 65 pure deictic; 729 total). These findings support *H1*.

It was surprising to see some participants using deictic instructions in *Side-by-Side*. I believe that they tried, failed, and switched to less unambiguous strategies such as using coordinates relative to the container where the disc was picked. It was also surprising that almost all participants pointed with their hands in all VIDEO conditions, even though they clearly knew that pointing would not be correctly understood in *Follow-Local* and *Side-by-Side*.

7.4.3.2 *Performer Error*

I investigate the role of VIDEO on *Performer*s producing errors (*PE*) when interpreting instructions, especially deictic ones. I remove instruction and interaction errors from the analysis, leaving 246 misunderstanding errors. Overall, participants produced fewer errors in *Follow-Remote* (66, 27% of total), followed by *Follow-Local* (82, 33% of total) and *Side-by-Side* (98, 40% of total).

Follow-Remote accounted for fewer errors if we consider the total number of deictic instructions produced in each VIDEO condition. 36% (4/11) of deictic instructions led to an error in *Follow-Local* and 40% (20/50) in *Side-by-Side*, but only 5% (21/410) in *Follow-Remote*. Deictic instructions were better interpreted in *Follow-Remote* than in the other VIDEO conditions, supporting *H2*.

7.4.3.3 *Word Count*

Word count (*WC*) can be used as a measure of communication efficiency. Using fewer words to communicate the same information suggests that the communication is more efficient, because the information is transmitted through other non-linguistic channels—video in our case. Participants used significantly different number of words in each VIDEO condition ($F_{2,3739} = 50.0747$, $p < 0.0001$). As expected, in *Follow-Remote*, *Instructor*s used significantly fewer words (2.98 ± 2.66 words) per instruction than in *Follow-Local* (3.80 ± 3.42 words) and *Side-by-Side* (4.07 ± 3.52 words).

I also investigate the number of deictic pronouns used by *Instructor*s. In *Follow-Remote*, 272 deictic pronouns were used, 110 in *Side-by-Side* and only 70 in *Follow-Local*.

In summary, when providing instructions in *Follow-Remote*, *Instructor*s used fewer words but more deictic pronouns than in other VIDEO conditions. To illustrate this point, *Instructor*s in *Follow-Local* typically used more verbose instructions, e.g., *"two to the left, then top"*, whereas in *Follow-Remote* they used short instructions with a deictic pronoun, e.g., *"top"* once they were in the correct column or simply *"there"* while pointing.

7.4.4 *Qualitative Analysis*

Participants answered a short questionnaire at the end of each VIDEO condition, and a final questionnaire at the end of the experiment. Questionnaires had both Likert scales and open questions.

The questionnaire for the different VIDEO conditions had two identical parts, one for each ROLE. All questions were in a five-level Likert scale[4], except for Q7 where the answer was on a continuous scale. Most questions were about perceived attention to each other:

- Q1: "I paid attention to my partner";
- Q2: "My partner paid attention to me";
- Q3: "It was easy to understand my partner";
- Q4: "My partner found it easy to understand me";
- Q5: "My behavior was in direct response to my partner's behavior";
- Q6: "The behavior of my partner was in direct response to my behavior";
- Q7: asked to estimate how much time they spent looking at the video when classifying objects (on a scale from 1 to 100); and,
- Q8: asked to assess how useful was the video of their partner for solving the task.

Data shows that for *Performer*s video was significantly more useful in *Follow-Remote* (mean 4.42) than in *Follow-Local* (mean 2.67, $p = 0.0057$) and *Side-by-Side* (mean 2.25, $p = 0.0012$). For *Instructor*s, video was significantly more useful in *Follow-Remote* (mean 2.83) than in *Side-by-Side* (mean 1.50, $p = 0.0.0372$). Also, *Instructor*s had the impression that their partner paid significantly more attention to them in *Follow-Remote* (mean 4.58) than in *Side-by-Side* (mean 3.83, $p = 0.0372$).

There is also an effect of VIDEO on how much participants looked at video both as *Performer* ($F_{2,22} = 13.24$, $p = 0.0002$) and *Instructor* ($F_{2,22} = 11.44$, $p = 0.0004$). *Performer*s used the video significantly more in *Follow-Remote* ($87 \pm 17\%$ of time) than in *Follow-Local* ($p = 0.0047$, $53 \pm 35\%$ of time) and *Side-by-Side* ($p = 0.0002$, $40 \pm 33\%$ of time). *Instructor*s used the video significantly more in *Follow-Remote* ($55 \pm 36\%$ of time) than in *Follow-Local* ($p = 0.023$, $30 \pm 25\%$ of time) and *Side-by-Side* ($p = 0.0003$, $15 \pm 22\%$ of time).

The final questionnaire asked participants (Q1) if they understood their partner's instructions when acting as *Performer* and (Q2) if their partner understood their instructions when acting as *Instructor*. It also asked (Q3-4-5) how often they used each *Instructor strategy* (*IS*) in each VIDEO condition, (Q6-7) their preferred VIDEO condition as *Instructor* and as *Performer*, and (Q8-9-10) a description of how they used the video in each VIDEO condition. All questions were likert scale from 1–5, except for preference where participants ranked the three condition in the order of their preference.

Participants reported that, as *Performer*s, they understood their partner's instructions significantly better in *Follow-Remote* (mean 4.63)

4 Likert-scale data are analyzed using Wilcoxon rank sum tests with Bonferroni correction.

than in *Follow-Local* (mean 4.00, $p = 0.0327$). Also, they indicated objects using instructions that are relative to the video (e.g. *"the one left of my video"*) significantly less often in *Side-by-Side* (mean 1.17) than in *Follow-Local* (mean 2.17, $p = 0.0354$) and *Follow-Remote* (mean 2.75, $p = 0.0009$). There were no other significant effects.

For evaluating participant preference, I assigned points according to the ranking (first equals 3 points, second 2 and third 1). *Instructor*s significantly preferred *Follow-Remote* (mean 2.29) over *Side-by-Side* (mean 1.58, $p = 0.0084$), and they significantly preferred *Follow-Local* (mean 2.16) over *Side-by-Side* ($p = 0.0378$). *Performer*s significantly preferred *Follow-Remote* (mean 2.29) over both *Side-by-Side* (mean 1.58, p's < 0.0003) and *Follow-Local* (mean 1.15, p's < 0.0003).

Video in *Follow-Remote* was used by *Performer*s to *"see where [the Instructor] was and then get the column where the object should be"* (P5) and *"to know on which column I have to place my object"* (P6). It also allowed them to *"follow [the Instructor's] position around the wall"* (P7). Many *Performer*s used the video *"to know what column [the Instructor] wants to pick and sometimes even the row"* (P9). I also observed that they used the video to determine gaze and predict the destination container more quickly: *"to get [the Instructor's] position, even the gaze helped me"* (P21). Some *Performer*s cleverly used the video in *Follow-Local* to estimate their partner's position. As people move, *CamRay* switches from one camera to the next in the array to capture their faces. These *Performer*s counted the "jumps" of the video window to *"roughly figure out how much I should move to the left/right"* (P23).

Surprisingly, only half the *Instructor*s ranked *Follow-Remote* first. *Follow-Local* was ranked first 10 times, and *Side-by-Side* 2 times. This was confirmed by participants when asked to describe how they used the video in *Follow-Local* and *Side-by-Side*: *"to see if [the Performer] was moving the object or not"* (P5), *"to know if my partner was focusing on the same task"* (P6), *"to get gaze direction and gestures, not position"* (P7) and *"to confirm verbal instructions"* (P20).

These findings partially support *H3*: almost all *Performer*s preferred *Follow-Remote* (22/24), but half of the *Instructor*s (12/24) preferred having the video in front of them or on the side. *Instructor*s liked seeing their remote collaborator as they performed instructions to check for understanding.

To summarize the results, in *Follow-Remote* participants:

- used more deictic instructions than in other VIDEO conditions, supporting *H1*;
- classified discs more efficiently, used fewer words and produced fewer misunderstanding errors, supporting *H2*; and,
- preferred this condition as *Performer*, but half did not prefer it as *Instructor*s, partially supporting *H3*.

7.5 DISCUSSION

The results provide evidence that the increased performance of *Follow-Remote* is related to (1) *Performer*s more closely following the *Instructor*s' position and gaze (*Cursor-Person Difference*, *Gaze-Position Difference*); and, (2) *Instructor*s using more deictic instructions (*IS*), leading to fewer *Performer error* (*PE*); and, (3) *Instructor*s using fewer words (*WC*).

First, *Performer*s were able to predict the target container as *Instructor*s moved and looked at the display: once an *Instructor* found a target container, the *Performer* would already be hovering a disc nearby and gazing in the vicinity, requiring less time to move and drop the disc. Second, as *Performer*s made fewer errors, they saved time. Third, awareness of the remote person's actions allowed for short and simple instructions, such as *"just there!"* or *"one above!"*.

Participants in this task had to *ground* their instructions, as the results can be explained by the natural tendency to minimize communication costs when generating common ground. Let us consider Clark's costs of grounding [26] in mediated communication for this experiment. Certain costs do not exist: there is no *start-up* time, and no *delay* nor *asynchrony* since communication was synchronous and real-time. Other costs are the same across VIDEO conditions: *production*, *reception* and *speaker change*, since all conditions used video-mediated communication; *fault* and *repair* since the severity of a fault and the time and effort to repair it dependended mainly on the task. We are thus left with three costs: *formulation*, *understanding* and *display*.

Formulation cost states that *"it costs more to plan complicated than simple utterances"* and *"to formulate perfect than imperfect utterances"* [26]. Different strategies have different costs: an instruction that relies on a coordinate system for absolute mapping, e.g., *"on container 3, 2"*, or a relative mapping to a container, e.g., *"two up, one down"* are costlier than pointing and using a pure deictic pronoun, e.g., *"there!"*. This suggests that *Follow-Remote* had a lower formulation cost.

Understanding cost states that *"the costs can be compounded when contextual clues are missing"* [26]. This explains why, when using deictic pronouns in *Follow-Local* or *Side-by-Side*, participants produced more errors: the cost of understanding is higher since the context to interpret instructions is missing.

Finally, *display* cost states that *"In media without copresence, gestures cost a lot, are severely limited, or are out of the question. In video teleconferencing, we can use only a limited range of gestures."* [26]. This explains why in *Follow-Remote*, *Instructor*s were able to use more deicitc gestures and these were understood more accurately by *Performer*s, reducing the display cost. This also explains why *Performer*s preferred *Follow-Remote* when interpreting instructions, while half the *Instructor*s preferred *Follow-Local* or *Side-by-Side*, since they could more easily check the *Performer*s for understanding.

In summary, by presenting video according to the remote collaborator's location in *Follow-Remote*, I enabled participants to use and

understand deictic instructions, reducing the overall cost of communication.

7.5.1 *Design Recommendations*

The experiment results and the analysis above lead to the following recommendations for the design of telepresence systems for wall-sized displays.

7.5.1.1 *Camera Arrays Support Remote Collaboration*

In large interactive spaces that allow physical navigation, an array of cameras mounted on a large display is an effective solution for remote collaboration. These cameras, when placed at eye's level can capture people's faces as they move, and this video can be displayed remotely. The camera setup is not limited to *CamRay*: we can envision adding other cameras to record people from different angles, even dynamically changing their own position, to support further needs during collaboration.

7.5.1.2 *Video Placed in Congruence with Content Better Supports Remote Pointing*

When collaborating across wall-sized displays, understanding deictic instructions depends on the accuracy of remote pointing interpretation. Displaying the remote participant's video in congruence with the shared space creates an instance of Buxton's Reference Space [21]. This experiment shows how collaborative data manipulation tasks in particular benefit from this setup, but there are many other tasks that also benefit from understanding deictic instructions.

7.6 SUMMARY AND CONTRIBUTIONS

In this chapter, I presented an experiment that uses *CamRay* to improve an asymmetric data manipulation task by supporting remote pointing. I leverage collaborators position by making video feeds follow the local person's position (*Follow-Local*) or the remote person's position (*Follow-Remote*), and compare it to a baseline condition where video does not move (*Side-by-Side*) to investigate how this affects collaboration. Participants were able to manipulate data more efficiently, taking less time, making fewer errors and using fewer words when video followed the remote collaborator. This can be explained by the fact that video enabled them to use and better understand deictic instructions, reducing the cost of communication.

Based on these findings, I provide recommendations for the design of future telecommunication systems across wall-sized displays: (1) camera arrays support remote collaboration, as they deal with user movement and (2) video placed in congruence with content better supports remote pointing.

Lastly, I found that some participants liked having video always visible, either in front of them or on the side, when checking their partner's understanding of instructions. This suggests that *Follow-Local* has benefits, which I explore in the next chapter.

8

SUPPORTING REPRESENTATIONAL GESTURES

This chapter reports on an experiment to validate the observation from Chapter 6 that virtual face-to-face might benefit the correct understanding of remote representational gestures. Pairs of participants perform a problem-solving task that heavily relies on hand gestures to represent objects. The results indicate that making the collaborator's video visible enables face-to-face communication at a distance, which I call a virtual face-to-face. This encourages the production of representational gestures and facilitates its understanding.

8.1 INTRODUCTION

In Clark et al.'s words [29], demonstrative references *"require an accompanying gesture for its complete interpretation"*. As I observed in Chapter 3, we can distinguish two types of communication during collaboration: gestural communication and conversational communication. Each of these is supported by the use of demonstrative references, but in a different way. In the previous chapter I showed how leveraging the remote collaborator's position to present video better supports remote *pointing gestures*, those used during gestural communication. The observations in Chapter 6 suggest that leveraging the position of the observer (*Follow-Local*) during face-to-face communication when talking about shared objects, might improve the remote counterpart's understanding. In this chapter, I investigate how to convey *representational gestures* at a distance, i.e. gestures that support conversational communication.

I use *CamRay* (Chapter 5) to evaluate the benefits of the *virtual face-to-face* (Figure 54) created when both collaborators are always visible to each other even when located at opposite sides of their respective displays (*Follow-Local*). I perform an experiment where two participants must exchange information using hand gestures to solve a problem. I present the rationale for the task and the types of stimuli used. I conclude by expanding the previous chapter guidelines for the design of telecommunication systems across wall-sized displays. This chapter is based on work submitted to the CHI 2018 conference in collaboration to Wendy Mackay.

8.2 GESTURES IN COMMUNICATION

The medium has an effect on remote collaboration [71], on aspects such as initiating conversation, establishing common ground and maintaining awareness of changes in the collaborative environment. Using video as a medium to convey gestures during remote work plays a

major role in communication. Bekker et al. [12] studied gestures in face-to-face design teams to provide guidelines for remote collaboration systems. They recognize that, among the many uses of gestures, these can be used to characterize concrete parts of a design, simulate its use and emphasize verbally stated points about the design. Kirk et al. [68] found that showing hand gestures in remote settings without changing their representation leads to faster performance with no loss of accuracy. This is partly due to grounding in communication [30] being supported by visual feedback, which enhances mutual awareness of actions and inter-subject intelligibility [67]. Campisi & Özyürek [23] used a teaching task to study the use of iconic gestures, a type of representational gestures. They compared strategies when teaching children, novices and expert adults. They found that the rate of iconic gestures increased, and they were more informative and bigger when directed towards children. The authors acknowledge that iconic gestures can be a powerful communicative strategy for teaching.

I dicussed in Chapter 2 the first systems that supported remote gestures: VideoDraw [106], VideoWhiteboard [107], Clearboard [61] and Commune [15]. All overlaid collaborators with shared drawings to support remote gestures. The authors not only observed that gestures help collaborators indicate parts of the shared content (pointing), but also that gestures help emphasize speech. This approach also increases awareness of remote participant's actions, as observed by Tang et al. [102] with tabletops when overlaying shadows of the remote collaborators' arms over the shared content, as they are more expressive than telepointers.

Previous work provides evidence that hand gestures play a crucial role in communication, in particular to represent objects and actions. I believe these benefit also apply to remote hand gestures that describe objects and actions during collaboration on shared data across wall-sized displays.

8.3 OPERATIONALIZING REPRESENTATIONAL GESTURES

My goal is to operationalize representational gestures and evaluate how a telecommunication system based on camera arrays, such as *CamRay*, supports this form of communication. I hypothesize that the virtual face-to-face created when the video of the remote collaborator is always in front of the local collaborator, even as they move (*Follow-Local*), supports representational gestures. In addition, I want to assess the trade-offs of both video behaviors for this type of gestures.

The task should reflect a realistic scenario where two experts need to combine their knowledge to solve a problem. Operationalizing such a task involves a number of constraints: The solution should be unique and objective, to avoid personal judgments and negotiations when choosing among possibilities; It should also be obtainable without the need to memorize or process information, to minimize

Figure 53: *Hands* and *Symbols* image sets.

the role of participants' memory or cognitive abilities. In short, the task should (1) require collaboration, (2) rely on the use of representational gestures, (3) have only one possible solution, (4) use only information that is on the display.

The task I created consists of describing an image to a remote participant. It is set up across two remote wall-sized displays, each one with one participant. The task is asymmetric, i.e. the participants have different roles. The *Instructor* is presented with an image on her display and has to describe it to the remote *Performer*, who has to search for it among a set of images presented on his display. To ensure that participants use hand gestures and not just words, I investigated through pilot studies how people describe several types of images to each other: tools, abstract symbols, people performing sign language, and people performing actions. I observed a heavier use of hand gestures, with some accompanying speech, when describing both sign language and abstract symbols. The participants found the sign language images easy and straightforward to describe, as they could replicate the hand pose. However, for the abstract symbols, participants used different types of descriptions combining gestures and speech. Based on these observations, I created two image sets for the task.

The *Hands* image set (Figure 53 left) features 99 pictures of a person performing hand poses inspired by sign language. I stood as an actor to perform the hand poses while watching a video of a news commentator translating voice into sign language, from [20]. The poses do not necessarily correspond to actual sign language words, as many were copied from the transitions between signs.

The *Symbols* image set (Figure 53 right) features 99 combinations of Khmer (Cambodian) characters. Consonants and vowels are combined with various subscripts and vowel diacritics, forming a set of characters composed of more than one part. Most of the produced characters are not valid in Khmer.

8.4 STUDYING REMOTE REPRESENTATIONAL GESTURES ACROSS WALL-SIZED DISPLAYS

I conducted an experiment across two wall-sized displays interconnected with *CamRay* to investigate the impact of video position on

collaboration for the task just described. The video of each collaborator is presented remotely in one of three ways:

1. *Follow-Local*: the video feed follows the position of the local collaborator, creating a virtual face-to-face with the remote collaborator;
2. *Follow-Remote*: the video feed follows the position of the remote collaborator; and,
3. *Side-by-Side* (control condition): the video feed is fixed on a separate screen on the side, perpendicular to the wall-sized display.

I use the two image sets described in the previous section in order to assess the effect of the video conditions on the use of representational gestures for both concrete and abstract stimuli: *Hands* are straightforward as users can imitate the pose; whereas for *Symbols* users must invent descriptive gestures.

I formulate five hypotheses with respect to the communication of representational gestures across wall-sized displays:

- *H1: Follow-Local is more efficient than Follow-Remote to perform the task;*
- *H2: Follow-Local leads to the production of more representational gestures than Follow-Remote;*
- *H3: Follow-Local is less intrusive for communication than Follow-Remote;*
- *H4: Follow-Local incurs a lower perceived workload than Follow-Remote; and,*
- *H5: the task relies more on representational gestures for Hands than for Symbols.*

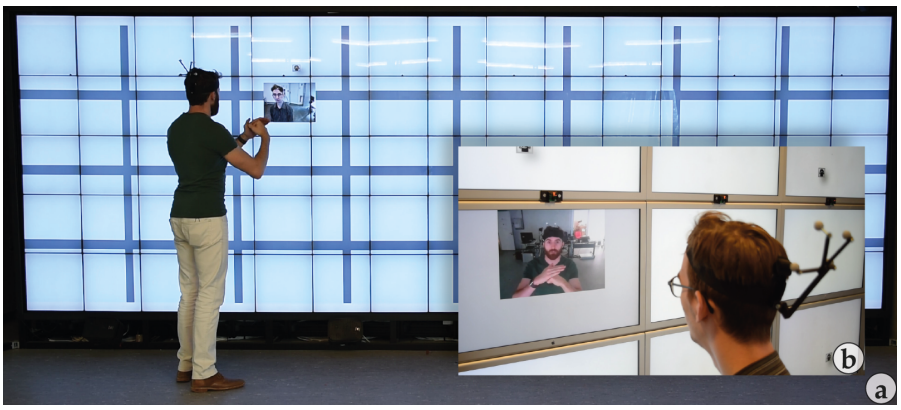


Figure 54: Experimental setup and virtual face-to-face. (a) A participant in *WILDER* describing an image using his hands. (b) A participant in *WILD* looking at the gesture.

8.4.1 *Method*

The experiment has a [3 × 2 × 2] within-participant design, with one secondary factor:

- VIDEO has three behaviors: *Follow-Remote*, *Follow-Local* & *Side-by-Side*;
- IMAGETYPE corresponds to the two image sets: *Hands* & *Symbols*; and,
- ROLE (secondary) to the asymmetric roles in the collaborative task: *Instructor* & *Performer*.

Participants perform the experiment in pairs, switching roles. The $3 \times 2 \times 2$ conditions are counterbalanced across participants using Balanced Latin squares. The order of the three VIDEO conditions are first counterbalanced, then for each condition the ROLES of the participants are switched. In each resulting block, a pair performs 3 replications of the *Hands* image set and 3 replications of the *Symbols* image set, alternating between the two.

I created 5 sets for each IMAGETYPE: 2 for training and 3 for the actual task. Each task set has 21 images: 15 distractors and 6 targets. For each VIDEO condition, I rotate through the 6 target images, ensuring that a different image is described in each trial.

8.4.2 *Participants*

12 participants (2 female), aged 21 to 31 (mean 25) participated in our study. None knew Khmer, and only one had vague notions of sign language. 8 collaborate daily with other people in the same space, of which 4 do so remotely as well. 6 use teleconferencing systems such as Skype on a weekly basis. Since the participants were from different nationalities, they could choose which language to use during the experiment. Two pairs chose to communicate in French, four in English.

8.4.3 *Hardware and Software*

As in the previous experiment, the setup is composed of the two wall-sized displays, *WILD* and *WILDER* (see Appendix A for details on these rooms), connected using *CamRay*. The array of 8 cameras embedded in the bezels capture participant's faces as they move; *CamRay* displays their video remotely using the local (*Follow-Local*) or remote (*Follow-Remote*) person's position. The positions of participants are tracked with a *VICON* motion tracking systems in each room. The videos are mirrored horizontally to preserve spatial references: a gesture performed by the right hand is seen remotely on the right hand side of the video.

Each room has an additional display and video camera, perpendicular to the wall-sized display, for the *Side-by-Side* condition: each

display shows the video captured by the other display, mirrored horizontally for consistency.

All video feeds are displayed at 34.7cm × 26cm (4:3 aspect ration), and presented at a fixed height of 1.75m. Audio is captured by wireless wearable microphones, sent through a *Google Hangout* call and played remotely using a high-quality surround sound system.

Interaction with both wall-sized displays uses touch input. *WILDER* has an infrared touch frame by PQLabs [1]. *WILD* does not support touch, so participants wear a wrist strap with markers tracked by the *VICON* system. The system registers a touch when the hand is close to the screen. Since this method is not as precise as a touch screen, the systems displays a cross at the position of the touch event so that participants can easily correct their hand position if needed. Note that precise pointing is not necessary for the task as participants only need to select images that measure 5cm × 4.2cm.

Webstrates [69] manages the content of and actions on the two wall-sized displays. The *Instructor*'s display shows the one image to be described to the *Performer*. The *Performer* sees the 16 images of the current set, one of which is the same as the one shown to the *Instructor*.

8.4.4 *Procedure*

Participants are welcomed and given written instructions about the experiment. They fill out an initial questionnaire and then are equipped with a head mount with reflective infrared markers and a wireless wearable microphone.

For each trial, I load the image together with a green button on the *Performer*'s display, and the target image covered by a gray rectangle on the *Instructor*'s display. Once the *Performer* touches the green button, the *Instructor* can touch the gray square to reveal the image and start the trial. This ensures that both participants are ready to start, and forces them to stand at the same position at the beginning of each trial.

The *Instructor* communicates with the *Performer* to describe the image with words and/or gestures. When the *Performer* thinks she has found the image described by the *Instructor*, she touches it. The system provides feedback on both sides by changing the selected image's border on the *Performer*'s side and the target image's border on the *Instructor*'s side. If the selection does not match the target image, both borders turn red for two seconds and another image can then be selected. If they match, their borders turn green. The trial ends when the *Performer* selects the matching target image. Participants are instructed to perform the task as quickly as possible and to make as few errors as possible.

When a dyad has performed all the trials for one VIDEO condition, participants fill out a short questionnaire. At the end of the experiment, they fill out a final questionnaire. Before each new VIDEO con-

1 http://www.pqlabs.com/

dition, participants perform 4 training trials (2 ROLE ~ 2 IMAGETYPE). The experiment takes about 70 minutes overall, 50 of which are spent performing the trials.

8.4.5 *Data Collection*

The system logs the layout of each image set as well as the time and location of touch events for the green button, the gray square and the image selections. The *VICON* motion tracking system records a kinematic log of the participants' position and orientation in each room. The sessions are video recorded in each room.

The initial questionnaire asks about past experience with large displays, remote collaboration, and knowledge of sign and Khmer languages. The intermediate questionnaires asks participants to assess the strengths and weaknesses of the VIDEO conditions, and how well they could understand each other. In the final questionnaire, participants are asked to describe their strategies to communicate and look for images in each VIDEO condition, how much they could understand each other, their preferred video setting for each ROLE, how they handled image occlusion by the video feeds, and to what extent the two IMAGETYPES led to different strategies.

8.4.6 *Data Analysis*

I analyze three types of data: performance data in terms of time and error, the kinematic logs to assess participants' motion, and the video recordings to analyze the communication patterns and the use of speech vs. gestures.

8.4.6.1 *Performance*

I measure the Task Completion Time (*TCT*) from the time where the *Instructor* has tapped the grey square to the time the *Performer* has found the correct target. I also count the number of errors for each trial, i.e. the number of incorrect selections, and the time from the beginning of the trial until the first selection, whether it is correct or not.

8.4.6.2 *Kinematic Analysis*

I observed that during remote conversations, participants wanted to see each other's faces. In *Follow-Local* participants moved less than in other conditions, as they could see each other without changing their positions. I therefore measure *Movement Quantity*, both for *Performer*s and *Instructor*s, to compute the amount of movement during communication.

Conversely, in *Follow-Remote*, participants moved to the remote collaborator's video when holding conversations. They synchronized their positions to see each other's video, standing "in front of each other" (in the same place relative to each one's room). I thus com-

pute *Position Synchronization*, to quantify how distant participants are from each other.

I normalize user position in each room so that we can compare them in both rooms, which have slightly different dimensions. For the horizontal axis, -1 corresponds to the left side of the screen, 0 to the center and 1 to the right.

8.4.6.3 *Video and Speech Analysis*

I transcribed video sessions with Chronoviz [41] and annotate and log the times of both speech and representational gestures.

Utterances are tagged using one or more categories:

- *Task*: working on the problem;
- *Technology*: overcoming technology issue;
- *Ask For Clarification*: asking to clarify a previous utterance;
- *Reply To Clarification*: replying to a clarification request;
- *Reference Previous*: using a reference to a previous trial; and
- *Common Ground*: using common ground, the shared knowledge, beliefs and assumptions between two people [28].

The number of task-related utterances reflects communication efficiency, while the number of technology-related utterances reflects how much the technology hinders collaboration. Analyzing clarification requests and replies can tell us if participants change strategy in case of a breakdown. Finally, I am interested in measuring common ground, as it can make conversation more efficient [29, 58], and reference to previous trials as this can indicate grounding [30].

For representational gestures, I identify gestures that are related to the task when describing images. For instance, a hand pose or drawing in the air with a finger. Within this set of gestures, I tag the following categories when appropriate:

- *Ask For Clarification*: asking to clarify a previous utterance,
- *Reply To Clarification*: replying to a clarification request; and
- *Copy Gesture*: copying a gesture while searching.

When asking for or replying to clarifications, if the person uses a gesture to accompany speech I tag it as *Ask For Clarification* or *Reply To Clarification* respectively. When searching for the target image, if the *Performer* uses his hands to "store" the *Instructor*'s gesture, I tag it as *Copy Gesture*.

8.5 RESULTS

A total of 269 task image touch events were registered overall (excluding practice trials), resulting in 216 trials (3 VIDEO $\times$ 2 IMAGETYPE $\times$

3 replications × 12 PARTICIPANT). I aggregate the 3 replications per condition using means, resulting in 72 data points. Unless otherwise specified, I perform full factorial analyses[2] of each measure for the model VIDEO × IMAGETYPE × Rand(PARTICIPANT) using REsidual Maximum Likelihood (REML) to account for random factors.

I removed 3 trials where participants had an unusual hard time solving the task (4 errors or more and time above 2 standard deviations). In order to test the hypotheses, I first analyze performance in terms of time and error, and then the quantity of movement of participants and the separation between their positions. I then analyze their use of speech and gestures, and finally qualitative data from the questionnaires[3].

8.5.1 *Task Performance*

A Shapiro-Wilk W test indicates that Task Completion Time (*TCT*) is not normally distributed ($W = 0.957$, $p = 0.0154$). A Kolmogorov's D test yields non-significant results ($D = 0.069$, $p = 0.15$), indicating that the data is log normal. Therefore, I work with the log transformation of *TCT* [88]. All analyses using the non-transformed time data yield the same effects with very similar p values.

The analysis of Task Completion Time *TCT* yields no significant main effect of VIDEO ($F_{2,55} = 0.0294$, $p = 0.971$), a main effect of IMAGETYPE ($F_{1,55} = 26.274$, $p < 0.0001$), and no VIDEO × IMAGETYPE interaction effect ($F_{2,55} = 0.992$, $p = 0.377$) (Figure 55). *Hands* images are solved faster than *Symbols* (32.86±11.30s vs. 45.71±13.92s).

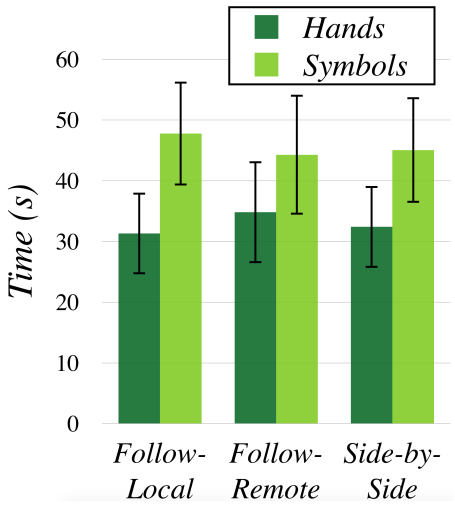


Figure 55: Time (*TCT*) in seconds for each VIDEO and IMAGETYPE condition. Bars show 95% confidence intervals.

From the total 269 image selections, 53 correspond to errors (i.e. $269 - 216 = 53$). I analyze the mean number of errors per trial. The analysis of *ERROR* yields no significant main effect of VIDEO ($F_{2,55} =$

2 All analysis are performed using SAS JMP

3 Questionnaire data are analyzed using Wilcoxon rank sum tests with Bonferroni correction.

VIDEO cond.	*Hands*	*Symbols*
Follow-Local	0.00	0.26
Follow-Remote	0.31	0.19
Side-by-Side	0.14	0.19

Table 3: Mean error (*ERROR*) per trial for each VIDEO and IMAGETYPE condition.

1.639, $p = 0.204$), nor of IMAGETYPE ($F_{1,55} = 1.610$, $p = 0.210$), but a VIDEO $\times$ IMAGETYPE interaction effect ($F_{2,55} = 3.928$, $p = 0.0254$) (Table 3). A post-hoc analysis[4] reveals that *Hands* in *Follow-Local* (ERROR = 0) has significantly fewer errors than *Hands* in *Follow-Remote* (0.31 ± 0.30 errors).

8.5.2 *Kinematic Analysis*

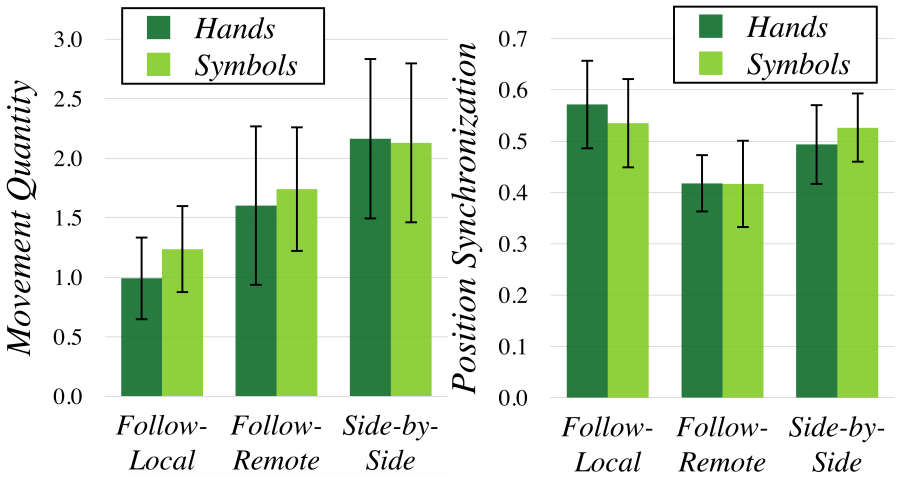
The kinematic log collected 237.477 data points overall representing the 3D positions of each participant. In order to measure the total distance traveled by participants and how much they synchronize their positions (how much they follow each other), I compute *Movement Quantity* by adding the distance between consecutive normalized positions of each participant along the X axis. Then, *Position Synchronization* is computed as the sum of the absolute differences between the normalized positions of the two collaborators along the X axis. These measures are averaged across each set of 3 replications, leading to 72 data points for each measure.

8.5.2.1 *Movement Quantity*

The analysis of *MQP* (*Movement Quantity Performer*) yields no significant main effect of VIDEO ($F_{2,55} = 2.4012$, $p = 0.100$), nor of IMAGETYPE ($F_{1,55} = 2.98$, $p = 0.0900$), nor of the VIDEO $\times$ IMAGETYPE interaction ($F_{2,55} = 1.7597$, $p = 0.1816$).

The analysis of *MQI* (*Movement Quantity Instructor*) yields a significant main effect of VIDEO ($F_{2,55} = 10.1987$, $p = 0.0002$), but not of IMAGETYPE ($F_{1,55} = 0.3898$, $p = 0.5350$), nor of the VIDEO $\times$ IMAGETYPE interaction ($F_{2,55} = 0.1904$, $p = 0.8272$) (Figure 56 a). A post-hoc test shows that in *Follow-Local* (1.12 ± 0.57) traveled distances are shorter than in *Follow-Remote* (1.67 ± 0.93) and *Side-by-Side* (2.15 ± 1.04).

4 All post-hoc analyses are performed using a Tukey-Kramer "Honestly Significant Difference" (HSD) test



(a) Movement Quantity of Ins. (*MQI*)

(b) Position Synchronization (*PS*)

Figure 56: Movement Quantity of Instructors (*MQI*) and Position Synchronization (*PS*) for each VIDEO and IMAGETYPE condition.

8.5.2.2 *Position Synchronization*

Position Synchronization (*PS*) reflects how much participants follow each other. The analysis of *PS* yields a significant main effect of VIDEO ($F_{2,55} = 10.684$, $p < 0.0001$), but not of IMAGETYPE ($F_{1,55} = 0.054$, $p = 0.820$) nor of the VIDEO $\times$ IMAGETYPE interaction ($F_{2,55} = 1.07$, $p = 0.350$) (Figure 56 b). A post-hoc test shows that participants are significantly closer to each other in *Follow-Remote* (0.38 ± 0.12) than in *Follow-Local* (0.54 ± 0.15) or *Side-by-Side* (0.48 ± 0.13).

8.5.3 *Video and Speech Analysis*

For measures of the video and speech analysis, I only report the analysis of categories that yield significant effects.

8.5.3.1 *Words*

Task-related words per trial are used significantly more in *Symbols* than in *Hands* (51.24 ± 25.5 vs. 21.11 ± 19.96 for *Instructors*, 12.79 ± 10.94 vs. 6.27 ± 7.86 for *Performers*, p's < 0.0002). This is consistent with our assumption that people would use more verbal descriptions in one set. The same trend also holds for the time spent talking.

Technology-related words per trial (e.g., *"Can you move?"* (P10)) are used significantly more in *Follow-Remote* than in both *Follow-Local* and *Side-by-Side* (2.79 ± 3.04 vs. 0 and 0.35 ± 0.83 for *Instructors*, p's < 0.0001; 2.90 ± 5.11 vs. 0.25 ± 0.78 and 0.39 ± 1.29 for *Performers*, p's < 0.012).

This effect is also significant for the time spent speaking: *Instructors* need to ask their partner to move when their video occludes an image, as the video position is controlled remotely, and the *Performer* acknowledges the request.

Clarifications are more often requested by *Performer*s with *Symbols* (0.99±0.71 per trial) than with *Hands* (0.57±0.75 per trial, $p = 0.0025$), but not differently across VIDEO conditions.

Common ground use is higher for *Symbols* than for *Hands* (2.38± 1.47 vs. 0.17±0.30 for *Instructor*s, $p < 0.0001$; 0.29±0.40 vs. 0.04±0.11 for *Performer*s, $p = 0.0003$). For *Symbols*, *Instructor*s use explanations such as *"The part above is like a mustache"* (P1). I was surprised to see the use of common ground for *Hands*, with expressions such as *"A Spok?"* (P2). I also noticed grounding [30], accruing common ground, as the experiment advances: *Instructor*s reference a previous trial to describe the current one more often for *Symbols* than for *Hands* (0.25± 0.32 vs. 0, $p < 0.0001$).

8.5.3.2 *Hand Gestures*

There were no significant differences for the overall gesture production time. I was surprised to observe however that *Instructor*s used hand gestures when replying to a clarification request significantly more in *Follow-Local* (35±32% of replies) than in *Follow-Remote* (9±16% of replies, $p = 0.0004$) or *Side-by-Side* (13±17% of replies, $p = 0.0033$).

The time spent gesturing is significantly longer for *Hands* than for *Symbols* (21.06±8.18s vs. 15.48±9.44s for *Instructor*s ($p = 0.0008$), 7.00±7.04s vs. 16.42±27.08s for *Performer*s ($p < 0.0001$)) Finally, *Performer*s copy gestures using their own hands while looking for solutions significantly more for *Hands* than for *Symbols* (4.65±5.18s vs. 0.01±0.07s ($p < 0.0001$)).

8.5.4 *Qualitative Analysis*

*Performer*s and *Instructor*s felt they understood their partner's instructions better in *Follow-Local* than in *Follow-Remote* (p's < 0.02).

I use the Expanded NASA-TLX [31] questionnaire to evaluate task load in each VIDEO condition. Participants found that *Follow-Local* was the least physically demanding condition (p's < 0.01) and that *Follow-Remote* was harder than *Follow-Local* ($p = 0.0132$), more frustrating ($p = 0.0387$), needed more coordination ($p = 0.03$) and it was harder to get support from the partner ($p = 0.0399$).

Regarding video occluding content, participants rated *Side-by-Side* as the condition with the least problems (p's < 0.0001) but, surprisingly, there were no significant differences between *Follow-Local* and *Follow-Remote*.

To evaluate participants' preference, I asked them to rank the VIDEO conditions. I assigned points according to the ranking (first equals 3 points, second 2 and third 1). *Instructor*s significantly preferred *Follow-Local* (2.75±0.45) over both *Follow-Remote* (1.16±0.38, $p < 0.0001$), and *Side-by-Side* (2.08±0.67, $p = 0.0393$). *Side-by-Side* was also preferred to *Follow-Remote* ($p = 0.0039$). *Performer*s significantly preferred *Follow-Local* (2.66±0.65) to *Follow-Remote* (1.41±0.51, $p = 0.0012$).

Nonetheless, in the open comments of the questionnaire, some participants mentioned that although they did not prefer *Follow-Remote*

for this task, they still felt it was more natural, as *Follow-Local* does not show how the other person moves even though it is clear that they are moving.

Finally, regarding the differences between the two image sets, participants felt that they share more previous knowledge with the *Symbols* than the *Hands* ($p = 0.0183$). They also report that the *Symbols* were harder to describe ($p = 0.0006$) and to understand ($p = 0.0004$) than the *Hands*.

8.6 DISCUSSION

These results show how the technology, but also the information being transmitted, affect the production and interpretation of *representational gestures*. I first summarize the results in terms of the initial hypotheses:

H1 (Follow-Local is more efficient than Follow-Remote to perform the task) is partially supported for the *Hands* image set: although there is no significant difference in time, participants made fewer errors with *Follow-Local* than with *Follow-Remote*.

H2 (Follow-Local leads to the production of more representational gestures than Follow-Remote) is partially supported: while the overall amount of gestures produced is not significantly different, when answering to clarification requests, *Instructor*s use more gestures for *Follow-Local* than for *Follow-Remote*.

H3 (Follow-Local is less intrusive for communication than Follow-Remote) is supported as both roles use fewer words with *Follow-Local* than with *Follow-Remote* to overcome technology problems, and they do not need to synchronize their movements.

H4 (Follow-Local incurs a lower perceived workload than Follow-Remote) is partially supported as 5 out of the 12 criteria were in favor of *Follow-Local*: physical demand, success as a team, frustration, need for coordination, need for communication and difficulty to have partner's support. The 7 other criteria were not significantly different across VIDEO conditions.

H5 (the task relies more on representational gestures for Hands than for Symbols) is supported as participants use more hand gestures and talk less for *Hands* than for *Symbols*.

8.6.1 *Image Sets*

As expected, results show that *Instructor*s spent more time talking and less time gesturing to describe *Symbols* than *Hands*. In particular, they used more common ground for describing *Symbols*. This is consistent with Clark & Wilkes-Gibbs's [30] *least collaborative effort* principle: people try to use as little combined effort as possible when grounding their conversations. We know from previous research that replacing nonverbal behaviors with verbal substitutes takes more time and effort [18]. This validates the choice of the two image sets: *Hands* required participants to use hand gestures, while *Symbols* were more

open and participants invented their own way to describe the images. As a consequence, *Symbols* were harder to describe as it took more time to finish the task and it required more clarifications.

8.6.2 *Errors*

Follow-Remote lead to more errors than *Follow-Local* for *Hands*, as describing these images relies heavily on representational gestures.

I believe this is due to what I call the *Video Avatar Effect*: throughout the experiment and in all VIDEO conditions, participants acted towards the remote video window as if it were their partner's eyes and ears. *Instructor*s wanted to talk directly to the video, so they either walked towards the video or gestured and talked towards it, which hindered communication. Even though participants were well aware that the system always captures their faces with the camera in front of them and their voices with the wearable microphone, they could not help but treat the video as an avatar of the remote collaborator. I even observed that some of them talked louder when the video was further away.

In *Follow-Remote*, this effect had two consequences that lead to more errors: First, when *Instructor*s walk to their partner's video feed, they leave the image to describe behind, leading to more errors as they rely on their memory and *"try to remember it to show it on the screen after"* (P2). Second, the *Instructor*'s hands were often outside the frame and *Performer*s interrupted them to let them know (*"I [only] see your head and torso"* (P6)). *Instructor*s also regularly asked to confirm visual contact (*"Can you see me?"* (P5)). Another common behavior was that *Instructor*s lowered their arms and stopped performing gestures as soon as the remote video started moving, as if their partners could not see them anymore.

On the contrary, *Follow-Local* was less hindered by the *Video Avatar Effect*: since the video feed stayed in front of the participants, *Instructor*s were accurately captured when they gestured or talked to it, and *Performer*s always had this information accessible. The experiment data supports this as participants found it easier in *Follow-Local* than in *Follow-Remote* to achieve the task and get support from their partners.

In the observations using *CamRay* (Section 6.4.5) I also observed that the video window was seen as a proxy for the remote collaborator. Some participants kept talking and gesturing towards empty videos, when the remote person was out of frame (e.g. when kneeling). I believe there is interest in further exploring how this effect can be leveraged to benefit communication.

8.6.3 *Representational Gestures*

The virtual face-to-face created by *Follow-Local* lets participants keep visual contact throughout the task. Although I did not observe differences in the production of gestures overall, when replying to clar-

ification requests, *Instructor*s used significantly more gestures with *Follow-Local* than with either *Follow-Remote* or *Side-by-Side*. I believe that this is because when they provided the first explanation, participants spontaneously created a face-to-face situation by synchronizing their positions. But when a clarification was required, participants were not always facing each other in *Follow-Remote* or *Side-by-Side*, meanwhile in *Follow-Local*.

What I did not expect was that *Performer*s would also use gestures so much for the *Hands* image set. I observed that participants oftentimes "stored" gestures using their own hands while moving and searching for the correct image. They used this as a way of externalizing knowledge instead of memorizing the hand poses, relying on Distributed Cognition [56]. With the *Symbols*, however, this process was harder as the images are more abstract. Some *Performer*s repeated parts of the instructions to themselves, which can be interpreted as a memorizing strategy.

8.6.4 *Technology Hindrance*

The experiment shows that the technology is less intrusive in *Follow-Local* than in the other conditions. Participants did not need to synchronize their positions and could solve occlusion by moving, as opposed to asking their partner. *Follow-Local* does not suffer from the issues of the other conditions: *Instructor*s traveled less than with *Side-by-Side* as moving to the side was not necessary to see the video, and participants did not have to synchronize their position as in *Follow-Remote* to create a face-to-face situation.

Conversely, *Follow-Remote* constantly "got on the way" of solving the task as it required participant to use more words related to technology than with the other conditions. Participants found *Follow-Remote* was more physically demanding, more frustrating and needed more coordination than *Follow-Local*.

8.6.5 *Preference*

The previous results about errors, support for representational gestures and technology hindrance can explain why *Follow-Local* leads to a lower perceived task workload, and why both *Instructor*s and *Performer*s ranked it as their preferred VIDEO condition.

8.6.6 *Design Recommendations*

Based on these results, I expand on the design recommendations from the previous chapter for telepresence systems across wall-sized displays.

8.6.6.1 *Video in the Focus Area Better Supports Representational Gestures*

Video in the focus area better supports representational gestures, both when producing them, as it does not require participants to synchronize their positions and encourages them to clarify instructions using gestures, and when interpreting them, as it leads to fewer errors. When video is far away, people tend to move and look for it even when it is not necessary to see it, hindering communication.

8.6.6.2 *Let Users Control Video Behavior*

Based on the results of the experiment in the previous chapter and this one, it becomes clear that different tasks require different video settings.. Users should therefore have some control for changing the behavior of video. Pointing tasks benefit from having the video follow the remote collaborator's position (*Follow-Remote*), but tasks that rely on representational gestures benefit from having the video follow the local collaborator's position (*Follow-Local*). Real tasks require both types of communication at different times, therefore systems should provide users with simple means to switch between these modes.

Sometimes, a task can require both types of communication at the same time, but on different sides. For example, in an teacher-learner scenario, the learner may want to see which objects the teacher is indicating (*Follow-Remote*), whereas the teacher may want to monitor the learner's reaction (*Follow-Local*). Therefore, systems should support asymmetric configurations. In addition, certain situations may benefit from one person controlling his remote representation. In this same scenario, the teacher might want to show a specific object to the learner after a virtual face-to-face, so he could enable *Follow-Remote* on the learner's side. Thus, systems might consider providing control over both local and remote video, so that users can control how they are presented remotely.

8.6.6.3 *Provide Local Mechanisms to Avoid Video Occlusion*

Although there are benefits to having video follow the remote collaborator for pointing tasks, it also has the disadvantage of generating unnecessary communication to deal with occlusion. Systems should therefore provide means to deal with occlusion locally. A solution is to let users explicitly control the position using, for example, a gesture to "push" the video aside, or to control the opacity of the video. This could be done e.g. by moving close to or away from the display, to see through the video. Another approach is to automatically adjust the position of the video according to the content of the display to minimize occlusion.

8.7 SUMMARY AND CONTRIBUTIONS

In this chapter I operationalize remote representational gestures in a collaborative problem-solving task across wall-sized displays using

concrete and abstract stimuli. I conducted and experiment using *CamRay* to explore how different video behaviors affect a task that relies on representational gestures.

I find that when the collaborator's video follows the local user, it creates a virtual face-to-face that encourages the production of gestures and makes their interpretation more accurate. For tasks that rely on representational gestures, people prefer this video behavior and perceive a lower task load compared to video that follows the remote user. These results complement the observations from Chapter 3 and Chapter 6, providing evidence of the benefits of leveraging the local user position.

I expand the design recommendations from the previous chapter for telecommunication systems across wall-sized displays, and recommend that (1) when video is shown in the focus area, it better supports representational gestures, (2) systems should provide users with ways to switch between video behaviors, and (3) systems should provide solutions to deal with video occluding content locally.

9

CONCLUSION

Wall sized displays introduce new opportunities for collaboration. They have the unique characteristic of presenting large data sets at a high resolution. This brings many benefits, from spatial tasks and sense-making tasks to search and comparison tasks. *Physical navigation* is at the heart of these benefits. People's ability to physically navigate data significantly outperforms panning and zooming, which we normally use instead on small displays.

Collaboration is of particular interest with wall-sized displays, as they can easily host small groups. Group members can comfortably interact with data and with each other, they can discuss information all together or form smaller groups and have side conversations. The synergy of two people working in collaboration comes from the combination of their unique and diverse skills. But as much as people enjoy working together on wall-sized displays, it is often the case that they are not co-located. We therefore need telecommunication systems to connect distant collaborators in these situations.

Already in the 1970's, Short [94] envisioned that a *"full-colour life-size three-dimensional motion-picture transmission of the complete body of the distant person by some futuristic television system"* cannot be the same as face-to-face contact as soon as people are aware that communication is mediated by technology. Hollan & Stronetta propose that systems must instead go *Beyond Being There* [57]. I promote their view in this thesis.

I argue that the success of a telecommunications system does not depend on its capacity to imitate co-located conditions, but on its ability to support the collaborative practices that emerge from the specific characteristics of the technology, and to create configurations that are not possible when co-located.

People choose collaborative technologies to support particular needs. When enabling the use of technology for remote collaboration, I believe that systems should support these needs, and leverage the characteristics of the technology.

This approach has been used in technologies other than wall-sized displays before. Tabletops for instance have a size, height and orientation that enable practices such as gathering around data and dividing the space into territories. Systems that implement remote collaboration through capturing and projecting video of remote arms convey the participants spatial location, which enables the formation of coupling styles at a distance. Territoriality on the other hand, can be overcome by these systems, as people do not interact with virtual arms in the same way as they do with physical ones.

In this thesis, I focus on wall-sized displays as a collaborative technology to explore how we can use the characteristics of the technology to support existing collaborative practices at a distance.

9.1 CONTRIBUTIONS

I start by empirically exploring how people might collaborate at a distance by simulating two remote locations on a wall-sized display (Chapter 3). Using simple prototypes, I explore what are the needs for collaboration in these spaces. I observe that people engage in two types of communication, *gestural communication* where they indicate shared digital objects by pointing at them, and *conversational communication* where they use hand gestures to support speech.

I investigate support for direct eye gaze and pointing gestures for wall-sized displays (Chapter 4). I find that people can accurately interpret remote pointing gestures when video of the person performing the gesture is congruent with the shared content. Also, the person and video can move without affecting both direct eye gaze perception and the accuracy of interpretation of remote pointing gestures.

I build a system for realizing remote collaboration across wall-sized displays called *CamRay* (Chapter 5). *CamRay* uses an array of cameras to capture people as they move in front of a display, and presents this video feed on a remote wall-sized display. The video can move across monitors controlled by different computers in a cluster, and it runs independently of the software on the display, allowing it to be used in a variety of applications.

CamRay can be used to explore the needs for remote collaboration across large interactive spaces, and how to support them. This is particularly useful, as wall-sized displays are hard to come across and real users for observations are not easy to find. The system has a flexible and open source architecture, which allows new cameras to be added and video feeds to be routed to new display surfaces. This provides facilities for prototyping new interaction techniques when envisioning novel telecommunication systems.

I observe how people use this tool (Chapter 6) and I propose two techniques that might support the two types of communication identified during my initial observations. These techniques leverage *physical navigation*, a common practice when exploring data on large displays. I call these two video behaviors *Follow-Local* and *Follow-Remote*. In *Follow-Remote*, the video feed of the remote participant follows the *remote* participant's movement. In *Follow-Local*, the video feed of the remote participant follows the *local* participant's movement.

Finally, I isolate critical aspects of the communication process based on my observations, and I conduct two experiments to explore how *CamRay* supports them. In the first one I explore *pointing gestures*, and in the second one *representational gestures*. In both cases I evaluate the use and understanding of gestures using *Follow-Local*, *Follow-Remote* and static video on the side. In the first experiment (Chapter 7), I adapt an existing data manipulation task so that it enforces remote collaboration. I find that when video follows the movement of the remote collaborator, it benefits the production and interpretation of pointing gestures. In a second experiment (Chapter 8), I design a task that relies on representational gestures, i.e. hand gestures to describe objects and the actions applied to them. I find that when video

follows the movement of the local person, it creates a *virtual face-to-face* that leads to better understanding of representational gestures. I show that *CamRay* can support existing activities in collaboration using wall-sized displays at a distance, by leveraging users' motion.

My original goal was to (1) keep the benefits of co-located collaboration in wall-sized displays, but in a remote setting; (2) overcome the limitations that arise from mediating communication with technology; and, (3) enable new configurations that are not possible in co-located settings.

First, many of the benefits of wall-size displays come from the fact that users physically navigate the space. *CamRay* aims at supporting physical navigation by displaying video that follows each collaborator, in order to keep these benefits. In this work, I performed studies for tasks that required manipulating or describing data. Further experiments need to be conducted to show that other collaborative tasks can be performed at a distance using *CamRay*.

Second, I show through experiments that we can overcome misunderstandings in pointing gestures by leveraging the position of the person producing such gestures. This approach preserves the spatial relation between people and content, creating a *reference space* [21]. Third, I show that we can enable remote, virtual face-to-face conversations by leveraging the position of the observers, which is a difficult configuration to achieve in a co-located setting with very wide displays.

9.2 IMPLICATIONS FOR DESIGN

I present in each chapter implications for the design of telecommunication systems across wall-sized displays. I now summarize these design guidelines.

Video feeds of remote collaborators can be moved on the wall-sized display without hindering the accuracy when interpreting remote indications, even when local collaborators move in the room. This gives flexibility to telecommunication systems when supporting physical movement, as they can move video on the display and do not need to limit users in their movements. This is of key importance for wall-sized displays, as the ability of users to physically move is at the core of its benefits. There are limitations on where to place the video feeds, which come from the two following implications.

Video Placed in Congruence With Content Better Supports Remote Pointing. To understand deictic instructions, video of the person performing the gesture should be placed in congruence with the object being indicated in the shared space. To achieve this, we can move the video of the remote collaborator performing the gesture to the correct position in the local space. Doing so lets the local collaborator to understand the gesture, without necessarily being in front of it.

Video Placed in the Focus Area Better Supports Representational Gestures. As collaborators move in the space, telepresence systems

can have the video of the remote participant follow them to keep each other's face within sight. This better supports representational gestures, both when producing and understanding them. Following the local user with the remote video creates a virtual face-to-face that allows collaborators to hold conversations and see each other even when they are on opposite ends of each one's rooms. This provides an advantage when the collaborators want to hold a discussion that involves pieces of shared content that are located in different parts of the display.

9.3 LIMITATIONS

Both the qualitative methods used when observing collaborators perform tasks using wall-sized displays, and the quantitative methods used when running controlled studies have their limitations.

Regarding the observations, the first limitation is not being able to perform them in a real environment with real users. Because wall-sized displays are not yet wildly adopted, it is hard to find regular users of this technology. As more people use wall-sized displays for collaboration in their daily work, observing them will provide insights of real problems and opportunities for improvement. Nonetheless, to increase the validity of my observations, I chose to observe people that were already collaborators, knew each other well, and were experts in the task.

Regarding the controlled studies, the second limitation is that they can be hard to generalize. On the one hand they provide hard evidence, but they can have limited ecological validity. As wall-sized displays become more widely adopted, it will be necessary to verify the results from the experiments in real-world settings.

In short, the low adoption of wall-sized displays as a technology limits (1) the possibility to conduct early observations that show problems and opportunities for improvement, and (2) the opportunity to assess the technology and give ecological validity to experimental results.

9.4 FUTURE WORK FOR WALL-SIZED DISPLAYS USING CAMRAY

This is but a first step in the exploration of how to support remote collaboration through video across wall-sized displays. First, the virtual face-to-face that *CamRay* enables should be further studied. In co-located collaboration, as wall-sized displays can grow to very large sizes, people may find themselves located far away from each other, even if in the same room. We can use *CamRay* to create a local virtual face-to-face and explore the question of when does being "far" becomes being "remote". In remote collaboration, we need to study how to engage in a virtual face-to-face with mutual agreement, avoiding abrupt transitions that might hinder collaboration. From the social perspective, there should be a clear protocol to start a virtual

face-to-face, just as we knock on the door when visiting someone's office.

Regarding the two video behaviors I studied, there is still the question of how to seamlessly integrate them. A dual user representation, one for position and pointing actions, and another one that conveys representational gestures, could potentially provide the benefits of both at the same time. This should be achieved without overloading users with dissociated representations that may be confusing.

For this integration, I envision that user movement and pointing gestures could use shadows that convey position and hand direction. This concept has been studied before, although not for collaboration, only to reach far away objects [93]. *CamRay*'s flexible architecture can accommodate more cameras for this purpose. For example, cameras on the ceiling could be used to detect collaborator's position and arm movement while pointing, and cameras on the back could be used to detect collaborators' silhouettes. This information can be overlaid over remote content and even on the floor to provide more ambient cues of remote pointing.

Another possibility for this integration is to move content instead of video feeds. Once two users engage in a virtual face-to-face, the video follows their own motion, which hinders the interpretation of deictic gestures. We can explore how having asymmetric content that adapts its layout to match pointing gestures can support remote indications.

Second, there is still a larger design space exploration that can be explored using *CamRay*. In this thesis, I take advantage of horizontal video movement, which is only one dimension. Moving video in the vertical dimension might also have an impact on communication. Geissner & Schubert [45] showed that vertical location has an impact on our judgment of leader's power. Is it possible that we can use these large displays to change, or even invert power relations? In this way, we would be using this technology for going beyond what's possible in a co-located situation.

The other dimension left to explore is the space in front of the wall. In this space we can use mobile telepresence robots to capture and display video, which could serve as simple forms of embodiment for remote collaborators, as shown in Figure 57.

Figure 57: A user in front of a wall-sized display collaborating with a remote person through a telepresence robot.

Analogous to the two video behaviors I studied, a mobile robot could always follow the remote person, reflecting where he is standing and looking, or always follow the local person, keeping a remote video feed at hand. We can even envision integrating video feeds from flying drones into *CamRay*, and display them either on the wall or on mobile screens that also move in the space. These drones can provide views of users from multiple angles as required by the task at hand.

Lastly, *CamRay* is designed to support multiple video feeds, allowing adding more users and more locations to the setup. From a technical perspective, we need to solve the challenge of selecting and displaying multiple video and audio feeds as more than two collaborators are present in more than two sites. From the perspective of collaboration, we need to explore the collaborative behaviors that occur in larger groups, such as coupling styles and territoriality.

9.5 COLLABORATION AT A DISTANCE FROM A BROADER PERSPECTIVE

This thesis shows an example of how to explore remote collaboration when incorporating it into existing technologies. Just because we can technologically achieve a goal, i.e. sending video across locations, does not mean that this is the best, or the only solution.

I observed how technology not only *mediates*, but also *modulates* communication at a distance. Showing video in different ways changed the way people communicate in my experiments. When collaborators need to perform an action, they use different strategies according to the available opportunities. They simply perform the action when the video allows them to, otherwise they adapt and turn to verbal descriptions, or move until the video position enables that action.

Technology affects behavior. People for example adapt their strategies when grounding conversation according to the cost of communication (Clark & Wilkes-Gibbs's [30] *least collaborative effort* principle). I believe that, in order to truly support remote collaboration, the exploration of how to integrate remote collaboration capabilities to collaborative technologies needs to focus on the existing collaborative practices, and how to support them at a distance by leveraging the characteristics of the technology.

Part I

APPENDIX

A

APPENDIX 1: WILD AND WILDER ROOMS

A.1 THE WILD ROOM

WILD (**W**all-sized **I**nteraction with **L**arge **D**atasets) consists of an 8×4 grid of 30" *Apple Cinema Display* screens (Figure 58). Each LCD monitor has a resolution of 2560×1600 pixes, and bezels of 30mm at the bottom and 22mm at the top, left and right. The entire wall-sized display measures 5.5m $\times$ 1.8m for a resolution of $20480 \times 6400 = 131,072,000$ pixels.

Figure 58: The *WILD* display.

WILD is controlled by a cluster of 16 *Apple Mac Pro* computers running an *Ubuntu* 14.04 *LTS* Linux distribution. Each computer manages two screens. The cluster is located in a room adjacent to the *WILD* room, and the computers are connected to the screens through long DVI cables associated with DVI repeaters. Three computers located in the *WILD* room, called *frontal*1–3, can be used independently to control the software running on the wall-sized display and distribute the rendering on the cluster. All the computers (*frontal*s and cluster) are connected to a local LAN network through the same router.

For interaction, a *VICON*[1] motion capture system allows to precisely determine the position and orientation of targets composed of several infrared reflective markers. It can be used to track both physical devices held by the user, enabling interaction with the display, and the users themselves, e.g. by wearing a cap, to follow their position in the room. Interactive devices (smartphones, tablets, tabletops, etc.)

1 https://www.vicon.com/

can connect to a dedicated WiFi network and be used for interacting with the platform. *WILD* also includes a 6.1 audio system.

A.2 THE WILDER ROOM

WILDER consists of a 15×5 grid of 21.6" *Planar*[2] screens. Each LCD monitor has a resolution of 960×960 pixes and bezels of 3mm (Figure 59). The entire wall-sized display measures 5.9m $\times$ 2m for a resolution of $14400 \times 4800 = 69,120,000$ pixels.

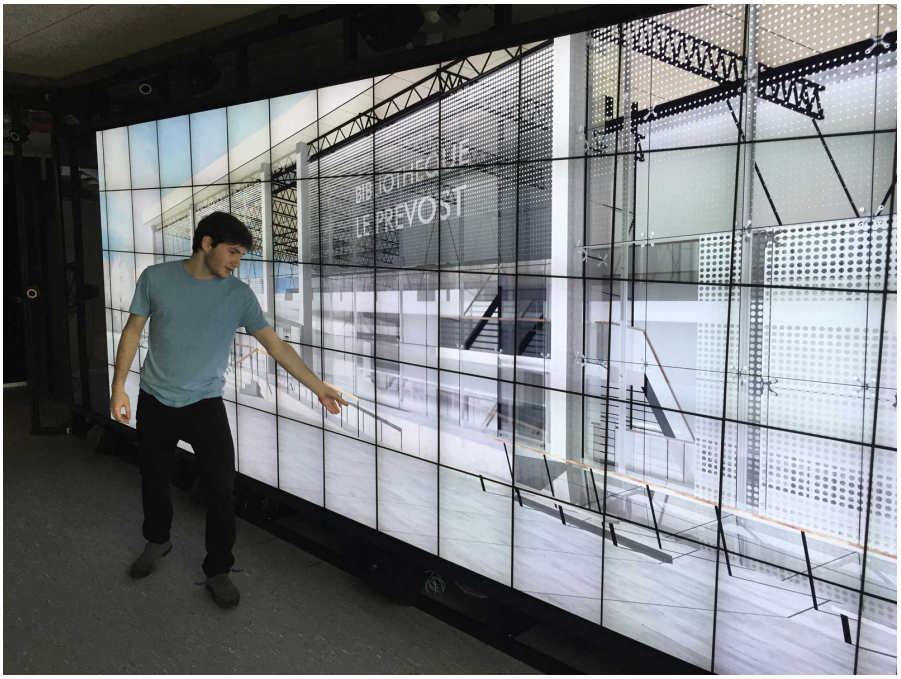


Figure 59: The *WILDER* display.

WILDER is controlled by a cluster of 10 *DELL* PCs an *Ubuntu* 14.04 *LTS* Linux distribution. Two computer manage an entire row of 15 screens: one for the first 8 monitors on the left and the other for the 7 remaining monitors on the right. The cluster is located in a server room in the basement of the building, and the computers are connected to the screens through optic fibers. Three computers located in the *WILDER* room, called *master*1–3, can be used independently to control the software running on the wall-sized display and distribute the rendering on the cluster. All the computers (*master*s and cluster) are connected on a dedicated LAN network across the building.

In addition to a *VICON* motion capture system, a *PQLabs*[3] infrared frame surrounding the wall-sized display adds multi-touch capability and enables several users to directly interact by touching the display. The room also includes a dedicated WiFi network to connect devices and a 8-speaker audio system.

2 http://www.planar.com/
3 http://www.pqlabs.com/

A.3 NETWORK CONNECTION

WILD and *WILDER* are connected through a local network. There is one computer in each room that receives the connections of the remote room. These two computers are connected through a virtual local network (VLAN) (Figure 60).

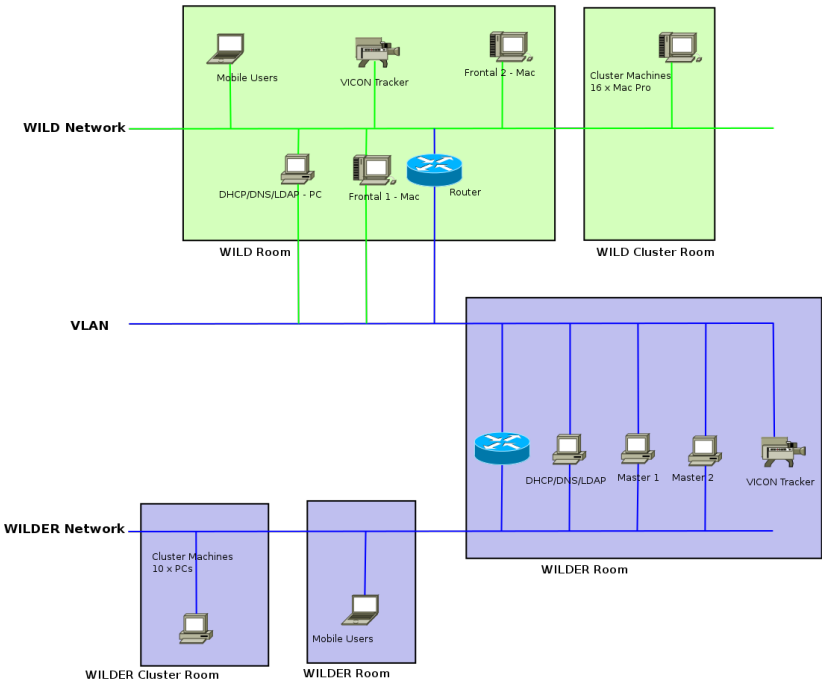


Figure 60: The *WILD* and *WILDER* room network architecture.

BIBLIOGRAPHY

[1] Deepak Akkil and Poika Isokoski. "Accuracy of Interpreting Pointing Gestures in Egocentric View." In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '16. New York, NY, USA: ACM, 2016, pp. 262–273. ISBN: 978-1-4503-4461-6. URL: https://doi.org/10.1145/2971648.2971687.

[2] Christopher Andrews, Alex Endert, and Chris North. "Space to Think: Large High-resolution Displays for Sensemaking." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, 2010, pp. 55–64. ISBN: 978-1-60558-929-9. URL: https://doi.org/10.1145/1753326.1753336.

[3] M. Argyle. *Social Interaction*. Methuen's manuals of modern psychology. Aldine, 1973. ISBN: 9780202368993.

[4] M Argyle. *Bodily Communication Routledge*. London, 1988.

[5] Ignacio Avellino, Cédric Fleury, and Michel Beaudouin-Lafon. "Accuracy of Deictic Gestures to Support Telepresence on Wall-sized Displays." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 2015, pp. 2393–2396. ISBN: 978-1-4503-3145-6. URL: https://doi.org/10.1145/2702123.2702448.

[6] Robert Ball and Chris North. "Analysis of User Behavior on High-resolution Tiled Displays." In: *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction*. INTERACT'05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 350–363. ISBN: 3-540-28943-7, 978-3-540-28943-2. URL: https://doi.org/10.1007/11555261_30.

[7] Robert Ball and Chris North. "Effects of Tiled High-resolution Display on Basic Visualization and Navigation Tasks." In: *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '05. New York, NY, USA: ACM, 2005, pp. 1196–1199. ISBN: 1-59593-002-7. URL: https://doi.org/10.1145/1056808.1056875.

[8] Robert Ball, Chris North, and Doug A. Bowman. "Move to Improve: Promoting Physical Navigation to Increase User Performance with Large Displays." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. New York, NY, USA: ACM, 2007, pp. 191–200. ISBN: 978-1-59593-593-9. URL: https://doi.org/10.1145/1240624.1240656.

[9] Adrian Bangerter. "Using Pointing and Describing to Achieve Joint Focus of Attention in Dialogue." In: *Psychological Science* 15.6 (2004), pp. 415–419. URL: https://doi.org/10.1111/j.0956-7976.2004.00694.x.

[10] Michel Beaudouin-Lafon et al. "Multisurface Interaction in the WILD Room." In: *Computer* 45.4 (Apr. 2012), pp. 48–56. ISSN: 0018-9162. URL: https://doi.org/10.1109/MC.2012.110.

[11] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. "Immersive Group-to-Group Telepresence." In: *IEEE Transactions on Visualization and Computer Graphics* 19.4 (Apr. 2013), pp. 616–625. ISSN: 1077-2626. URL: https://doi.org/10.1109/TVCG.2013.33.

[12] Mathilde M. Bekker, Judith S. Olson, and Gary M. Olson. "Analysis of Gestures in Face-to-face Design Teams Provides Guidance for How to Use Groupware in Design." In: *Proceedings of the 1st Conference on Designing Interactive Systems: Processes, Practices, Methods, & Techniques*. DIS '95. New York, NY, USA: ACM, 1995, pp. 157–166. ISBN: 0-89791-673-5. URL: https://doi.org/10.1145/225434.225452.

[13] A. Bezerianos and P. Isenberg. "Perception of Visual Variables on Tiled Wall-Sized Displays for Information Visualization Applications." In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2516–2525. ISSN: 1077-2626. URL: https://doi.org/10.1109/TVCG.2012.251.

[14] Xiaojun Bi and Ravin Balakrishnan. "Comparing Usage of a Large High-resolution Display to Single or Dual Desktop Displays for Daily Work." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: ACM, 2009, pp. 1005–1014. ISBN: 978-1-60558-246-7. URL: https://doi.org/10.1145/1518701.1518855.

[15] S. A. Bly and S. L. Minneman. "Commune: A Shared Drawing Surface." In: *Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems*. COCS '90. New York, NY, USA: ACM, 1990, pp. 184–192. ISBN: 0-89791-358-2. URL: https://doi.org/10.1145/91474.91514.

[16] Lauren Bradel, Alex Endert, Kristen Koch, Christopher Andrews, and Chris North. "Large High Resolution Displays for Co-located Collaborative Sensemaking: Display Usage and Territoriality." In: *Int. J. Hum.-Comput. Stud.* 71.11 (Nov. 2013), pp. 1078–1088. ISSN: 1071-5819. URL: https://doi.org/10.1016/j.ijhcs.2013.07.004.

[17] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology." In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. URL: https://doi.org/10.1191/1478088706qp063oa.

[18] Susan Elise Brennan. *Seeking and providing evidence for mutual understanding*. Stanford University, 1990.

[19] Harry Brignull, Shahram Izadi, Geraldine Fitzpatrick, Yvonne Rogers, and Tom Rodden. "The Introduction of a Shared Interactive Surface into a Communal Space." In: *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. CSCW '04. New York, NY, USA: ACM, 2004, pp. 49–58. ISBN: 1-58113-810-5. URL: https://doi.org/10.1145/1031607.1031616.

[20] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. "Long Term Arm and Hand Tracking for Continuous Sign Language TV Broadcasts." In: *Proceedings of the British Machine Vision Conference*. 2008.

[21] Bill Buxton. "Mediaspace – Meaningspace – Meetingspace." In: *Media Space 20 + Years of Mediated Life*. Ed. by Steve Harrison. London: Springer London, 2009, pp. 217–231. ISBN: 978-1-84882-483-6. URL: https://doi.org/10.1007/978-1-84882-483-6_13.

[22] William A. S. Buxton. "Telepresence: Integrating Shared Task and Person Spaces." In: *Proceedings of the Conference on Graphics Interface '92*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992, pp. 123–129. ISBN: 0-9695338-1-0. URL: https://dl.acm.org/citation.cfm?id=155294.155309.

[23] Emanuela Campisi and Asli Özyürek. "Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children." In: *Journal of Pragmatics* 47.1 (2013), pp. 14–27. DOI: https://doi.org/10.1016/j.pragma.2012.12.007.

[24] Milton Chen. "Leveraging the Asymmetric Sensitivity of Eye Contact for Videoconference." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '02. New York, NY, USA: ACM, 2002, pp. 49–56. ISBN: 1-58113-453-3. URL: https://doi.org/10.1145/503376.503386.

[25] Colin Cherry. *On Human Communication*. Studies in communication. Wiley, 1957.

[26] Herbert H. Clark and Susan E. Brennan. "Grounding in communication." In: *Perspectives on socially shared cognition*. Ed. by L. B. Resnick, J. M. Levine, and S. D. Teasley. Washington, DC, US: American Psychological Association, 1991, pp. 127–149. ISBN: 978-1-55798-121-9.

[27] Herbert H Clark and Meredyth A Krych. "Speaking while monitoring addressees for understanding." In: *Journal of memory and language* 50.1 (2004), pp. 62–81.

[28] Herbert H Clark and Catherine R Marshall. "Definite reference and mutual knowledge." In: (1981).

[29] Herbert H. Clark, Robert Schreuder, and Samuel Buttrick. "Common ground at the understanding of demonstrative reference." In: *Journal of Verbal Learning and Verbal Behavior* 22.2 (1983), pp. 245 –258. ISSN: 0022-5371. URL: https://doi.org/10.1016/S0022-5371(83)90189-5.

[30] Herbert H. Clark and Deanna Wilkes-Gibbs. "Referring as a collaborative process." In: *Cognition* 22.1 (1986), pp. 1 –39. ISSN: 0010-0277. URL: https://doi.org/10.1016/0010-0277(86)90010-7.

[31] Denis Coelho, João N. O. Filipe, Mário Simões-Marques, and Isabel Nunes. "The Expanded Cognitive Task Load Index (NASA-TLX) applied to Team Decision-Making in Emergency Preparedness Simulation." In: *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference*. 2014, pp. 225–236.

[32] Mary Czerwinski, Greg Smith, Tim Regan, Brian Meyers, George Robertson, and Gary Starkweather. "Toward Characterizing the Productivity Benefits of Very Large Displays." In: *PROC. INTERACT*. Press, 2003, pp. 9–16.

[33] Alan Dix, Janet Finlay, Gregory Abowd, and Russell Beale. *Human–Computer Interaction*. Prentice Hall, 1998. ISBN: 978-0130461094.

[34] M. Dou, Y. Shi, J. M. Frahm, H. Fuchs, B. Mauchly, and M. Marathe. "Room-sized informal telepresence system." In: *2012 IEEE Virtual Reality Workshops (VRW)*. Mar. 2012, pp. 15–18. URL: https://doi.org/10.1109/VR.2012.6180869.

[35] Paul Dourish and Victoria Bellotti. "Awareness and Coordination in Shared Workspaces." In: *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work*. CSCW '92. New York, NY, USA: ACM, 1992, pp. 107–114. ISBN: 0-89791-542-9. URL: https://doi.org/10.1145/143457.143468.

[36] H. J. Dudfield, C. Macklin, R. Fearnley, A. Simpson, and P. Hall. "Big is better? Human factors issues of large screen displays with military command teams." In: *2001 People in Control. The Second International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*. 2001, pp. 304–309. URL: https://doi.org/10.1049/cp:20010480.

[37] D. Efron. *Gesture and Environment: A Tentative Study of Some of the Spatio-temporal and Linguistic Aspects of the Gestural Behavior of Eastern Jews and Southern Italians in New York City, Living Under Similar as Well as Different Environmental Conditions*. Col. uni. diss. King's Crown Press, 1941. URL: https://doi.org/10.1177/000271624222000197.

[38] Paul Ekman and Wallace V Friesen. "The repertoire of nonverbal behavior: Categories, origins, usage, and coding." In: *semiotica* 1.1 (1969), pp. 49–98.

[39] Scott Elrod et al. "Liveboard: A Large Interactive Display Supporting Group Meetings, Presentations, and Remote Collaboration." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '92. New York, NY, USA: ACM, 1992, pp. 599–607. ISBN: 0-89791-513-5. URL: https://doi.org/10.1145/142750.143052.

[40] Alex Endert, Christopher Andrews, Yueh Hua Lee, and Chris North. "Visual Encodings That Support Physical Navigation on Large Displays." In: *Proceedings of Graphics Interface 2011*. GI '11. School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada: Canadian Human-Computer Communications Society, 2011, pp. 103–110. ISBN: 978-1-4503-0693-5. URL: https://dl.acm.org/citation.cfm?id=1992917.1992935.

[41] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D. Hollan. "ChronoViz: A System for Supporting Navigation of Time-coded Data." In: *CHI '11 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '11. New York, NY, USA: ACM, 2011, pp. 299–304. ISBN: 978-1-4503-0268-5. URL: https://doi.org/10.1145/1979742.1979706.

[42] Susan R. Fussell, Leslie D. Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam D. I. Kramer. "Gestures over Video Streams to Support Remote Collaboration on Physical Tasks." In: *Hum.-Comput. Interact.* 19.3 (Sept. 2004), pp. 273–309. ISSN: 0737-0024. URL: https://doi.org/10.1207/s15327051hci1903_3.

[43] Darren Gergle, Robert E. Kraut, and Susan R. Fussell. "Action As Language in a Shared Visual Space." In: *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. CSCW '04. New York, NY, USA: ACM, 2004, pp. 487–496. ISBN: 1-58113-810-5. URL: https://doi.org/10.1145/1031607.1031687.

[44] James J. Gibson and Anne D. Pick. "Perception of Another Person's Looking Behavior." In: *The American Journal of Psychology* 76.3 (1963), pp. 386–394. URL: https://doi.org/10.2307/1419779.

[45] Steffen R. Giessner and Thomas W. Schubert. "High in the hierarchy: How vertical location and judgments of leaders' power are interrelated." In: *Organizational Behavior and Human Decision Processes*. 2007, pp. 30–44. URL: https://doi.org/10.1016/j.obhdp.2006.10.001.

[46] Zenzi M. Griffin and Kathryn Bock. "What the Eyes Say About Speaking." In: *Psychological Science* 11.4 (2000), pp. 274–279. URL: https://doi.org/10.1111/1467-9280.00255.

[47] Jonathan Grudin. "Partitioning Digital Worlds: Focal and Peripheral Awareness in Multiple Monitor Use." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. New York, NY, USA: ACM, 2001, pp. 458–465.

ISBN: 1-58113-327-8. URL: https://doi.org/10.1145/365024.365312.

[48] François Guimbretière, Maureen Stone, and Terry Winograd. "Fluid Interaction with High-resolution Wall-size Displays." In: *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*. UIST '01. New York, NY, USA: ACM, 2001, pp. 21–30. ISBN: 1-58113-438-X. URL: https://doi.org/10.1145/502348.502353.

[49] Carl Gutwin and Saul Greenberg. "A Descriptive Framework of Workspace Awareness for Real-Time Groupware." In: *Computer Supported Cooperative Work (CSCW)* 11.3 (Sept. 2002), pp. 411–446. ISSN: 1573-7551. DOI: 10.1023/A:1021271517844. URL: https://doi.org/10.1023/A:1021271517844.

[50] Carl Gutwin, Saul Greenberg, and Mark Roseman. "Workspace Awareness in Real-Time Distributed Groupware: Framework, Widgets, and Evaluation." In: *People and Computers XI: Proceedings of HCI'96*. Ed. by Martina Angela Sasse, R. Jim Cunningham, and Russel L. Winder. London: Springer London, 1996, pp. 281–298. ISBN: 978-1-4471-3588-3. URL: https://doi.org/10.1007/978-1-4471-3588-3_18.

[51] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. "Spatiality in Videoconferencing: Trade-offs Between Efficiency and Social Presence." In: *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*. CSCW '06. New York, NY, USA: ACM, 2006, pp. 413–422. ISBN: 978-1-59593-249-5. URL: https://doi.org/10.1145/1180875.1180937.

[52] Kirstie Hawkey, Melanie Kellar, Derek Reilly, Tara Whalen, and Kori M. Inkpen. "The Proximity Factor: Impact of Distance on Co-located Collaboration." In: *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*. GROUP '05. New York, NY, USA: ACM, 2005, pp. 31–40. ISBN: 1-59593-223-2. URL: https://doi.org/10.1145/1099203.1099209.

[53] Christian Heath and Paul Luff. "Collaborative Activity and Technological Design: Task Coordination in London Underground Control Rooms." In: *Proceedings of the Second Conference on European Conference on Computer-Supported Cooperative Work*. ECSCW'91. Norwell, MA, USA: Kluwer Academic Publishers, 1991, pp. 65–80. ISBN: 0-7923-1439-5. URL: https://dl.acm.org/citation.cfm?id=1241910.1241915.

[54] D.Y.P. Henriques and J. D. Crawford. "Role of Eye, Head, and Shoulder Geometry in the Planning of Accurate Arm Movements." In: *Journal of Neurophysiology* 87.4 (2002), pp. 1677–1685. ISSN: 0022-3077. DOI: 10.1152/jn.00509.2001. URL: https://doi.org/10.1152/jn.00509.2001.

[55] Keita Higuchi, Yinpeng Chen, Philip A. Chou, Zhengyou Zhang, and Zicheng Liu. "ImmerseBoard: Immersive Telepresence Experience Using a Digital Whiteboard." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 2015, pp. 2383–2392. ISBN: 978-1-4503-3145-6. URL: https://doi.org/10.1145/2702123.2702160.

[56] James Hollan, Edwin Hutchins, and David Kirsh. "Distributed Cognition: Toward a New Foundation for Human-computer Interaction Research." In: *ACM Trans. Comput.-Hum. Interact.* 7.2 (June 2000), pp. 174–196. ISSN: 1073-0516. URL: https://doi.org/10.1145/353485.353487.

[57] Jim Hollan and Scott Stornetta. "Beyond Being There." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '92. New York, NY, USA: ACM, 1992, pp. 119–125. ISBN: 0-89791-513-5. URL: https://doi.org/10.1145/142750.142769.

[58] Ellen A Isaacs and Herbert H Clark. "References in conversation between experts and novices." In: *Journal of experimental psychology: general* 116.1 (1987), p. 26. DOI: http://dx.doi.org/10.1037/0096-3445.116.1.26.

[59] Ellen A. Isaacs and John C. Tang. "What Video Can and Can't Do for Collaboration: A Case Study." In: *Proceedings of the First ACM International Conference on Multimedia*. MULTIMEDIA '93. New York, NY, USA: ACM, 1993, pp. 199–206. ISBN: 0-89791-596-8. URL: https://doi.org/10.1145/166266.166289.

[60] Petra Isenberg, Danyel Fisher, Sharoda A. Paul, Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. "Co-Located Collaborative Visual Analytics Around a Tabletop Display." In: *IEEE Transactions on Visualization and Computer Graphics* 18.5 (May 2012), pp. 689–702. ISSN: 1077-2626. URL: https://doi.org/10.1109/TVCG.2011.287.

[61] Hiroshi Ishii and Minoru Kobayashi. "ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '92. New York, NY, USA: ACM, 1992, pp. 525–532. ISBN: 0-89791-513-5. URL: https://doi.org/10.1145/142750.142977.

[62] Mikkel R. Jakobsen and Kasper Hornbæk. "Up Close and Personal: Collaborative Work on a High-resolution Multitouch Wall Display." In: *ACM Trans. Comput.-Hum. Interact.* 21.2 (Feb. 2014), 11:1–11:34. ISSN: 1073-0516. URL: https://doi.org/10.1145/2576099.

[63] Mikkel R. Jakobsen, Yonas Sahlemariam Haile, Søren Knudsen, and Kasper Hornbæk. "Information Visualization and Proxemics: Design Opportunities and Empirical Findings." In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec.

2013), pp. 2386–2395. ISSN: 1077-2626. URL: https://doi.org/10.1109/TVCG.2013.166.

[64] Mikkel Jakobsen and Kasper Hornbæk. "Proximity and Physical Navigation in Collaborative Work with a Multi-touch Wall-display." In: *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '12. New York, NY, USA: ACM, 2012, pp. 2519–2524. ISBN: 978-1-4503-1016-1. URL: https://doi.org/10.1145/2212776.2223829.

[65] Adam Kendon. "Some relationships between body motion and speech." In: *Studies in dyadic communication* 7.177 (1972), p. 90.

[66] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[67] David Kirk, Andy Crabtree, and Tom Rodden. "Ways of the Hands." In: *Proceedings of the Ninth Conference on European Conference on Computer Supported Cooperative Work*. ECSCW'05. New York, NY, USA: Springer-Verlag New York, Inc., 2005, pp. 1–21. ISBN: 978-1402040221. URL: https://dl.acm.org/citation.cfm?id=1242029.1242030.

[68] David Kirk and Danae Stanton Fraser. "Comparing Remote Gesture Technologies for Supporting Collaborative Physical Tasks." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. New York, NY, USA: ACM, 2006, pp. 1191–1200. ISBN: 1-59593-372-7. URL: https://doi.org/10.1145/1124772.1124951.

[69] Clemens N. Klokmose, James R. Eagan, Siemen Baader, Wendy Mackay, and Michel Beaudouin-Lafon. "Webstrates: Shareable Dynamic Media." In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*. UIST '15. New York, NY, USA: ACM, 2015, pp. 280–290. ISBN: 978-1-4503-3779-3. URL: https://doi.org/10.1145/2807442.2807446.

[70] Robert M. Krauss, Robert A. Dushay, Yihsiu Chen, and Frances Rauscher. "The Communicative Value of Conversational Hand Gesture." en. In: *Journal of Experimental Social Psychology* 31.6 (Nov. 1995), pp. 533–552. ISSN: 00221031. DOI: 10.1006/jesp.1995.1024. URL: http://doi.org/10.1006/jesp.1995.1024 (visited on 10/10/2017).

[71] Robert E Kraut, Susan R Fussell, Susan E Brennan, and Jane Siegel. "Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work." In: *Distributed work* (2002), pp. 137–162.

[72] Russell Kruger, Sheelagh Carpendale, Stacey D. Scott, and Saul Greenberg. "Roles of Orientation in Tabletop Collaboration: Comprehension, Coordination and Communication." In: *Comput. Supported Coop. Work* 13.5-6 (Dec. 2004), pp. 501–537. ISSN: 0925-9724. URL: https://doi.org/10.1007/s10606-004-5062-8.

[73] M. Kuechler and A Kunz. "HoloPort - A device for simultaneous video and data conferencing featuring gaze awareness." In: *Proc. Virtual Reality, VR'06*. IEEE, 2006, pp. 81–88. ISBN: 1087-8270. URL: https://doi.org/10.1109/VR.2006.71.

[74] Lars Lischke, Sven Mayer, Katrin Wolf, Niels Henze, Albrecht Schmidt, Svenja Leifert, and Harald Reiterer. "Using Space: Effect of Display Size on Users' Search Performance." In: *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '15. New York, NY, USA: ACM, 2015, pp. 1845–1850. ISBN: 978-1-4503-3146-3. URL: https://doi.org/10.1145/2702613.2732845.

[75] Can Liu, Olivier Chapuis, Michel Beaudouin-Lafon, Eric Lecolinet, and Wendy E. Mackay. "Effects of Display Size and Navigation Type on a Classification Task." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: ACM, 2014, pp. 4147–4156. ISBN: 978-1-4503-2473-1. URL: https://doi.org/10.1145/2556288.2557020.

[76] Can Liu, Olivier Chapuis, Michel Beaudouin-Lafon, and Eric Lecolinet. "Shared Interaction on a Wall-Sized Display in a Data Manipulation Task." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. New York, NY, USA: ACM, 2016, pp. 2075–2086. ISBN: 978-1-4503-3362-7. URL: https://doi.org/10.1145/2858036.2858039.

[77] Andrés Lucero, Aaron Quigley, Jun Rekimoto, Anne Roudaut, Martin Porcheron, and Marcos Serrano. "Interaction Techniques for Mobile Collocation." In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. MobileHCI '16. New York, NY, USA: ACM, 2016, pp. 1117–1120. ISBN: 978-1-4503-4413-5. URL: https://doi.org/10.1145/2957265.2962651.

[78] Paul K. Luff, Naomi Yamashita, Hideaki Kuzuoka, and Christian Heath. "Flexible Ecologies And Incongruent Locations." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, NY, USA: ACM, 2015, pp. 877–886. ISBN: 978-1-4503-3145-6. URL: https://doi.org/10.1145/2702123.2702286.

[79] Andrew Maimone and Henry Fuchs. "Encumbrance-free Telepresence System with Real-time 3D Capture and Display Using Commodity Depth Cameras." In: *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*. ISMAR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 137–146. ISBN: 978-1-4577-2183-0. URL: https://doi.org/10.1109/ISMAR.2011.6092379.

[80] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[81] Andrew F. Monk and Caroline Gale. "A Look Is Worth a Thousand Words: Full Gaze Awareness in Video-Mediated Conversation." In: *Discourse Processes* 33.3 (2002), pp. 257–278. URL: https://doi.org/10.1207/S15326950DP3303_4.

[82] David Nguyen and John Canny. "MultiView: Spatially Faithful Group Video Conferencing." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '05. New York, NY, USA: ACM, 2005, pp. 799–808. ISBN: 1-58113-998-5. URL: https://doi.org/10.1145/1054972.1055084.

[83] Tao Ni, Doug A. Bowman, and Jian Chen. "Increased Display Size and Resolution Improve Task Performance in Information-Rich Virtual Environments." In: *Proceedings of Graphics Interface 2006*. GI '06. Toronto, Ont., Canada, Canada: Canadian Information Processing Society, 2006, pp. 139–146. ISBN: 1-56881-308-2. URL: https://dl.acm.org/citation.cfm?id=1143079.1143102.

[84] Daniele S. Pagani and Wendy E. Mackay. "Bring Media Spaces into the Real World." In: *Proceedings of the Third Conference on European Conference on Computer-Supported Cooperative Work*. ECSCW'93. Norwell, MA, USA: Kluwer Academic Publishers, 1993, pp. 341–356. ISBN: 0-7923-2447-1. URL: https://dl.acm.org/citation.cfm?id=1241934.1241957.

[85] Emmanuel Pietriga, Stéphane Huot, Mathieu Nancel, and Romain Primet. "Rapid Development of User Interfaces on Cluster-driven Wall Displays with jBricks." In: *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. EICS '11. New York, NY, USA: ACM, 2011, pp. 185–190. ISBN: 978-1-4503-0670-6. URL: https://doi.org/10.1145/1996461.1996518.

[86] Bordia Prashant. "Face-to-Face Versus Computer-Mediated Communication: A Synthesis of the Experimental Literature." In: *The Journal of Business Communication (1973)* 34.1 (1997), pp. 99–118. URL: https://doi.org/10.1177/002194369703400106.

[87] Margaret Gwendoline Riseborough. "Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication." In: *Journal of Nonverbal Behavior* 5.3 (Mar. 1981), pp. 172–183. ISSN: 1573-3653. URL: https://doi.org/10.1007/BF00986134.

[88] Judy Robertson and Maurits Kaptein. *Modern Statistical Methods for HCI*. Springer, 2016. ISBN: 978-3-319-26633-6. URL: https://doi.org/10.1007/978-3-319-26633-6.

[89] Stacey D. Scott, M. Sheelagh T. Carpendale, and Kori M. Inkpen. "Territoriality in Collaborative Tabletop Workspaces." In: *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. CSCW '04. New York, NY, USA: ACM, 2004, pp. 294–303. ISBN: 1-58113-810-5. URL: https://doi.org/10.1145/1031607.1031655.

[90] L.D. Segal. "Designing Team Workstations: The Choreography of Teamwork." In: *Local applications of the Ecological Approach to Human-Machine Systems, Vol. 2* (1995). Ed. by J. Caird & K. Vicente P. Hancock J. Flach.

[91] Abigail J. Sellen. "Remote Conversations: The Effects of Mediating Talk with Technology." In: *Human-Computer Interaction* 10.4 (Dec. 1995), pp. 401–444. ISSN: 0737-0024. URL: https://doi.org/10.1207/s15327051hci1004_2.

[92] Abigail Sellen, Bill Buxton, and John Arnott. "Using Spatial Cues to Improve Videoconferencing." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '92. New York, NY, USA: ACM, 1992, pp. 651–652. ISBN: 0-89791-513-5. URL: https://doi.org/10.1145/142750.143070.

[93] Garth Shoemaker, Anthony Tang, and Kellogg S. Booth. "Shadow Reaching: A New Perspective on Interaction for Large Displays." In: *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*. UIST '07. New York, NY, USA: ACM, 2007, pp. 53–56. ISBN: 978-1-59593-679-0. URL: https://doi.org/10.1145/1294211.1294221.

[94] John Short, Ederyn Williams, and Bruce Christie. *The social psychology of telecommunications*. John Wiley and Sons Ltd, 1976.

[95] Norbert A. Streitz, Jörg Geissler, Torsten Holmer, Shinichi Konomi, Christian Müller-Tomfelde, Wolfgang Reischl, Petra Rexroth, Peter Seitz, and Ralf Steinmetz. "i-LAND: An Interactive Landscape for Creativity and Innovation." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '99. New York, NY, USA: ACM, 1999, pp. 120–127. ISBN: 0-201-48559-1. URL: https://doi.org/10.1145/302979.303010.

[96] Desney S. Tan, Mary Czerwinski, and George Robertson. "Women Go with the (Optical) Flow." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. New York, NY, USA: ACM, 2003, pp. 209–215. ISBN: 1-58113-630-7. URL: https://doi.org/10.1145/642611.642649.

[97] Desney S. Tan, Darren Gergle, Peter Scupelli, and Randy Pausch. "With Similar Visual Angles, Larger Displays Improve Spatial Performance." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. New York, NY, USA: ACM, 2003, pp. 217–224. ISBN: 1-58113-630-7. URL: https://doi.org/10.1145/642611.642650.

[98] Desney S. Tan, Darren Gergle, Peter G. Scupelli, and Randy Pausch. "Physically Large Displays Improve Path Integration in 3D Virtual Navigation Tasks." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. New York, NY, USA: ACM, 2004, pp. 439–446. ISBN: 1-58113-702-8. URL: https://doi.org/10.1145/985692.985748.

[99] Desney S. Tan, Darren Gergle, Peter Scupelli, and Randy Pausch. "Physically Large Displays Improve Performance on Spatial Tasks." In: *ACM Trans. Comput.-Hum. Interact.* 13.1 (Mar. 2006), pp. 71–99. ISSN: 1073-0516. URL: https://doi.org/10.1145/1143518.1143521.

[100] Kar-Han Tan, I Robinson, R. Samadani, Bowon Lee, D. Gelb, A Vorbau, B. Culbertson, and J. Apostolopoulos. "ConnectBoard: A remote collaboration system that supports gaze-aware interaction and sharing." In: *IEEE International Workshop on Multimedia Signal Processing, 2009. MMSP '09*. 2009, pp. 1–6. ISBN: 978-1-4244-4463-2. URL: https://doi.org/10.1109/MMSP.2009.5293268.

[101] Anthony Tang, Michael Boyle, and Saul Greenberg. "Understanding and mitigating display and presence disparity in mixed presence groupware." In: *Journal of Research and Practice in Information Technology* 37.2 (2005), pp. 193–210. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.4331.

[102] Anthony Tang, Carman Neustaedter, and Saul Greenberg. "VideoArms: Embodiments for Mixed Presence Groupware." en. In: *SpringerLink*. Springer London, 2006, pp. 85–102. ISBN: 978-1-84628-664-3. URL: https://doi.org/10.1007/978-1-84628-664-3_8.

[103] Anthony Tang, Melanie Tory, Barry Po, Petra Neumann, and Sheelagh Carpendale. "Collaborative Coupling over Tabletop Displays." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. New York, NY, USA: ACM, 2006, pp. 1181–1190. ISBN: 1-59593-372-7. URL: https://doi.org/10.1145/1124772.1124950.

[104] Anthony Tang, Michel Pahud, Kori Inkpen, Hrvoje Benko, John C. Tang, and Bill Buxton. "Three's Company: Understanding Communication Channels in Three-way Distributed Collaboration." In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. CSCW '10. New York, NY, USA: ACM, 2010, pp. 271–280. ISBN: 978-1-60558-795-0. URL: https://doi.org/10.1145/1718918.1718969.

[105] John C. Tang. "Findings from observational studies of collaborative work." In: *International Journal of Man-machine studies* 34.2 (1991), pp. 143–160. URL: https://doi.org/10.1016/0020-7373(91)90039-A.

[106] John C. Tang and Scott L. Minneman. "VideoDraw: A Video Interface for Collaborative Drawing." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '90. New York, NY, USA: ACM, 1990, pp. 313–320. ISBN: 0-201-50932-6. URL: https://doi.org/10.1145/97243.97302.

[107] John C. Tang and Scott Minneman. "VideoWhiteboard: Video Shadows to Support Remote Collaboration." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '91. New York, NY, USA: ACM, 1991, pp. 315–322. ISBN: 0-89791-383-3. URL: https://doi.org/10.1145/108844.108932.

[108] Paul Tanner and Varnali Shah. "Improving Remote Collaboration Through Side-by-side Telepresence." In: *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '10. New York, NY, USA: ACM, 2010, pp. 3493–3498. ISBN: 978-1-60558-930-5. URL: https://doi.org/10.1145/1753846.1754007.

[109] D. G. Tatar, G. Foster, and D. G. Bobrow. "Computer-supported Cooperative Work and Groupware." In: ed. by Saul Greenberg. London, UK, UK: Academic Press Ltd., 1991. Chap. Design for Conversation: Lessons from Cognoter, pp. 55–80. ISBN: 0-12-299220-2. URL: https://doi.org/10.1145/259963.260448.

[110] Philip Tuddenham and Peter Robinson. "Territorial Coordination and Workspace Awareness in Remote Tabletop Collaboration." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: ACM, 2009, pp. 2139–2148. ISBN: 978-1-60558-246-7. URL: https://doi.org/10.1145/1518701.1519026.

[111] Elizabeth S. Veinott, Judith Olson, Gary M. Olson, and Xiaolan Fu. "Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '99. New York, NY, USA: ACM, 1999, pp. 302–309. ISBN: 0-201-48559-1. URL: https://doi.org/10.1145/302979.303067.

[112] Roel Vertegaal, Gerrit van der Veer, and Harro Vonsfect. "Effects of Gaze on Multiparty Mediated Communication." In: *Proceedings of Graphics Interface*. Morgan Kaufmann, 2000, pp. 95–102.

[113] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. "GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-controlled Camera Direction." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. New York, NY, USA: ACM, 2003, pp. 521–528. ISBN: 1-58113-630-7. URL: http://doi.org/10.1145/642611.642702.

[114] M. Von Cranach and J. Heinrich Ellgring. "The Perception of Looking Behaviour." In: *Social Commun. and Movement* (1973). Ed. by M. Von Cranach and Vine I.

[115] Malte Willert, Stephan Ohl, Anke Lehmann, and Oliver Staadt. "The Extended Window Metaphor for Large High-resolution Displays." In: *Proceedings of the 16th Eurographics Conference on Virtual Environments &; Second Joint Virtual Reality*. EGVE - JVRC'10. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2010, pp. 69–76. ISBN: 978-3-905674-30-9. URL: https://doi.org/10.2312/EGVE/JVRC10/069-076.

[116] Nelson Wong and Carl Gutwin. "Where Are You Pointing?: The Accuracy of Deictic Pointing in CVEs." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, 2010, pp. 1029–1038. ISBN: 978-1-60558-929-9. URL: https://doi.org/10.1145/1753326.1753480.

[117] Beth Yost, Yonca Haciahmetoglu, and Chris North. "Beyond Visual Acuity: The Perceptual Scalability of Information Visualizations for Large Displays." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. New York, NY, USA: ACM, 2007, pp. 101–110. ISBN: 978-1-59593-593-9. URL: https://doi.org/10.1145/1240624.1240639.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Communication médiatisée par la vidéo pour les pratiques collaboratives à distance entre murs d'écrans

Mots clés : murs d'écrans, communication distante, communication gestuelle, rangée de caméras

Résumé : La collaboration entre plusieurs personnes peut prendre plusieurs formes, et la technologie soutient depuis longtemps ces pratiques. Mais lorsque la collaboration doit se faire à distance, est-elle aussi bien assistée par la technologie ? Dans ce travail, je soutiens l'idée selon laquelle le succès d'un système de télécommunications ne dépend pas de sa capacité à imiter une collaboration colocalisée, mais dans sa capacité à faciliter les pratiques collaboratives découlant des caractéristiques spécifiques de la technologie. J'explore cet argument en utilisant un mur d'écrans en tant que technologie collaborative. J'ai commencé par observer des collaborateurs effectuer leur travail quotidien à distance en utilisant des prototypes. Ensuite j'ai conduit des expériences et j'ai trouvé que les utilisateurs peuvent interpréter avec précision les instructions déictiques à distance et le regard direct quand un collaborateur à distance est affiché par une vidéo, même si celle-ci n'est pas placée directement devant l'observateur. À partir de ces résultats, j'ai créé CamRay, un outil de télécommunication qui utilise une rangée de caméras pour enregistrer le visage des utilisateurs lorsqu'ils parcourent physiquement les données le long de l'écran et présente cette vidéo sur un autre mur d'écrans distant par dessus le contenu existant. Je propose deux possibilités pour afficher la vidéo: Follow-Local, où le flux vidéo de l'utilisateur distant suit l'utilisateur local, et Follow-Remote où il suit l'utilisateur distant. Je montre que Follow-Remote préserve les relations spatiales entre le collaborateur à distance et le contenu de l'écran, créant ainsi la possibilité de désigner les objets par des gestes de pointage, tandis que Follow-Local facilite les conversations grâce à un face-à-face virtuel qui transmet plus facilement la communication gestuelle. Finalement, je me base sur ces résultats pour guider la conception de futurs systèmes de communications à distance entre murs d'écrans, et dégager des considérations à suivre lorsque des capacités de communication à distance sont ajoutées à de nouvelles technologies.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Supporting Collaborative Practices Across Wall-Sized Displays with Video-Mediated Communication

Keywords: wall-sized displays, remote collaboration, gestures, camera array

Abstract: Collaboration can take many forms, for which technology has long provided digital support. But when collaborators are located remotely, to what extent does technology support these activities? In this dissertation, I argue that the success of a telecommunications system does not depend on its capacity to imitate co-located conditions, but in its ability to support the collaborative practices that emerge from the specific characteristics of the technology. I explore this using wall-sized displays as a collaborative technology. I started by observing collaborators perform their daily work at a distance using prototypes. I then conducted experiments and found that people can accurately interpret remote deictic instructions and direct gaze when performed by a remote collaborator through video, even when this video is not placed directly in front of the observer. Based on these findings, I built CamRay, a telecommunication system that uses an array of cameras to capture users' faces as they physically navigate data on a wall-sized display, and presents this video in a remote display on top of existing content. I propose two ways of displaying video: Follow-Local, where the video feed of the remote collaborator follows the local user, and Follow-Remote, where it follows the remote user.
I find that Follow-Remote preserves the spatial relations between the remote speaker and the content, supporting pointing gestures, while Follow-Local enables virtual face-to-face conversations, supporting representational gestures. Finally, I summarize these findings to inform the design of future systems for remote collaboration across wall-sized displays

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France