Mechanics of Camera Work in Mobile Video Collaboration

Brennan Jones[1], Anna Witcraft[1], Scott Bateman[2], Carman Neustaedter[3], Anthony Tang[1]

[1]University of Calgary
{bdgjones, anna.witcraft, tonyt}@ucalgary.ca

[2]University of Prince Edward Island
sbateman@upei.ca

[3]Simon Fraser University
carman@sfu.ca

ABSTRACT

Mobile video conferencing, where one or more participants are moving about in the real world, enables entirely new interaction scenarios (e.g., asking for help to construct or repair an object, or showing a physical location). While we have a good understanding of the challenges of video conferencing in office or home environments, we do not fully understand the *mechanics* of camera work—how people use mobile devices to communicate with one another—during mobile video calls. To provide an understanding of what people do in mobile video collaboration, we conducted an observational study where pairs of participants completed tasks using a mobile video conferencing system. Our analysis suggests that people use the camera view deliberately to support their interactions—for example, to convey a message or to ask questions—but the limited field of view, and the lack of camera control can make it a frustrating experience.

Author Keywords

CSCW; collaboration; handheld devices; mobile computing; video communication

ACM Classification Keywords

H5.3. Information interfaces and presentation: Group and Organization Interfaces – *CSCW*

INTRODUCTION

Recent advances in mobile technologies and networks are enabling new scenarios of video conferencing use—in particular, scenarios where one or more participants of a video call are out "in the real world" physically moving around (e.g., [12, 28]). For example, people now use mobile video conferencing to give tours of new places (e.g., [1, 20, 23]), share views of the outdoors or life experiences (e.g., [12]), provide directions to someone who is unfamiliar with a setting, help make decisions in retail outlets, and guide others during simple repair or construction tasks (e.g., [1]).

Mobile video conferencing differs from other video conferencing situations for two reasons: first, participants are away from a desk or a controlled environment (e.g., indoors, [12, 28]); second, that the video is being captured from a non-stable, moving camera. This presents an interesting paradox: on the one hand, manipulating the camera provides more control over the camera view; on the other hand, people must also actively attend to controlling the view, and the needs of the other party (e.g., [23]).

Figure 1: A mobile video conferencing scenario where one participant is out and about and the other is at a PC.

While others have explored the emerging social challenges and difficulties associated with mobile video conferencing (e.g., [26, 28]), our research is focused on the details of how people try to accomplish tasks using mobile video. Specifically, we are interested in the *mechanics of mobile camera work*—how people manipulate the handheld phone camera's position, orientation, the framing of objects, scenes, etc.—as a means of supporting collaborative interaction. For example: how do people help another person gain a sense of their spatial environment, and the objects in it? And, does this facilitate navigation? The overarching goal of our research is to articulate these mechanics, and challenges with existing technology, as a means of informing the design of new tools that support mobile video conferencing.

To help identify and articulate these mechanics, we designed and conducted a study where pairs of participants (connected via a mobile video call) were engaged in collaborative activities. To complete tasks, one partner used a mobile phone while 'out and about' to connect with their partner at a remote computer (Figure 1). We explore how the mobile device provides support for, and sometimes hinder, their actions and intentions through a variety of representative collaborative scenarios (e.g., [1, 11, 23, 26]).

Our results show that while video can help interaction by taking the place of or by supporting conversation, people compensate using a number of practices that are not ideal. In particular, camera work is hindered by the need for both

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea

ACM 978-1-4503-3145-6/15/04…$15.00
http://dx.doi.org/10.1145/2702123.2702345

good "overview" and "detail" views—often simultaneously. The imbalance of camera control means that one partner is often left frustrated, and without the view they need to support collaboration. Furthermore, because deixis is such an important part of communication, the lack of gesturing mechanisms can be problematic. Based on these results, we suggest ways to provide overviews, awareness, and gesturing within mobile video calls, as well as tools to negotiate control of the camera view.

We make two main contributions in this paper: first, we provide the first detailed study of mobile camera work, articulating the exact mechanics of how people try to communicate using mobile video in common collaborative scenarios; second, we outline several issues that people face in these scenarios that can inform the design of future mobile video conferencing technologies.

BACKGROUND

Portable, Personal Video Conferencing in the Home

Research on personal uses of video conferencing has explored the use of fixed-location and handheld devices (e.g., [1, 14, 15, 20, 34]). Several researchers have described how "open connections" facilitate the sharing of life and routine in the home (e.g., [15, 20]). We have also seen novel systems address specific uses of video conferencing in domestic settings ranging from the use of physical proxies [32] to interactive storybooks [29] to media spaces [14, 15]. These explorations have focused primarily on video conferencing in the home using relatively stable camera views. In contrast, our work explores the use video conferencing in mobile contexts—where one participant can move around freely.

Mobile Video and Shared Experiences

Mobile video conferencing has not yet been well-explored. O'Hara's [26] diary study explored why people made mobile video calls, and the challenges they faced. In his sample, 28% of such calls were to show objects in a scene, while 22% of calls were for functional needs (e.g., asking for or providing assistance). Challenges included technical and environmental problems (e.g., ambient noise, poor lighting, etc.), and social challenges (e.g., embarrassing to use in public). Licoppe et al. [23] analyzed 100 mobile video calls from eight pairs of users, and discuss the use of mobile video as an interactional resource for conversation (e.g., pointing the camera at an object helps support that conversation). Brubaker et al. [1] provide vignettes based on the real life use of mobile video (e.g., providing help to perform an oil change or repairing airplanes), and describe how its use in establishing relationships with distant places (e.g., providing "house tours").

Recent work has explored novel mobile video conferencing prototypes where participants are "out and about" in the real world. Inkpen et al. [12] explore the use of video to share children's outdoor activities with a remote person at home. Here, multiple views (one of the scene, the other of the local parent) were beneficial to remote participants of the video chat: they could observe the activities while seeing reactions of the local parent. Procyk et al. [28] explore remote geocaching where pairs of people participate while streaming video to each other. The authors note the value of "micro shared experiences," but also discuss privacy concerns and dangers of being overly engaged with a remote person while navigating outdoors. Another type of activity is the 'virtual photo walk,' where a person shares video of an experience (e.g., a hike) with people who are unable to take part [33]. While commercial mobile video chat systems (e.g., FaceTime and Skype) are now widely available, few studies of their use exist.

While these explorations provide us with insight into the possibilities of mobile video calls (e.g., [12, 28]) and the social challenges of public use (e.g., [23, 26]), we are specifically interested in understanding the ways that people use mobile devices, the interactional challenges they face, and how they cope with these by using camera work.

Remote gesture in Collaborative Tasks

A principal challenge identified in explorations of video conferencing systems is being able to effectively refer to objects in the video scene (e.g., [10, 27]). So-called "deictic references" (e.g., pointing at an object and saying, "this one") are made problematic because each party in a video call does not typically share "physical co-presence" [28].

Gergle et al. [10] explore the role of visual connections that focus on a workspace (such as a desk), showing how being able to see the workspace supports interaction by taking the place of speech. Here, the visual space acts as a resource for interaction—there is no need to confirm that a particular action was taken (e.g., putting a LEGO block in place) when all participants easily view the action, and it is possible to refer to objects through deictic reference [21]. Kirk et al. [18] demonstrate that when the workspace that is captured includes a person's hands working in the space, people use their hands in surprisingly expressive ways to communicate. Gutwin and Penner [11] demonstrate that there is a temporal element to gestures, and that they are fleeting. Displaying gestures with visual traces that fade away slowly helps interpretation over remote connections [6]. Sodhi et al. [31] explore the problem of object referencing in video conferencing and add augmented reality. Their approach uses a symmetric mobile tool that captures a remote participant's gestures and hand postures and represents them using an avatar. While this seems like a promising solution, it is not yet widely available.

Although research has widely recognized the importance of gesturing, and provided some innovative solutions to facilitate gesturing over video, video conferencing systems still do not provide much support to facilitate gesturing.

Free-Moving Video in Collaborative Tasks

To ease remote gesturing and enable a greater range of expressive interaction and sharing, research has also explored novel video configurations. Fussell et al. [3]

examine the combined use of head mounted and workspace-focused video for collaborative tasks. They demonstrate that head-mounted video provides a more detailed "work area," but has a limited FOV, and requires the helper to reorient himself as the worker moves head positions. In contrast, the scene camera provides the helper with a better sense of context, but a poorer sense of the specific activities being undertaken. Norris et al. [25] use a focus+context approach to give collaborators "focus windows" that provide detail in addition to the context window of the entire space. Yarosh et al. [34] explore the use of a mobile video camera condition in their study of child free-play, finding that movement and framing of children and scenes can be challenging.

We have a deep understanding of how video conferencing can support collaborative activities in work contexts where cameras are usually in fixed positions (e.g., [19]). Yet, we know comparatively little about mobile scenarios, where people need to manipulate the camera view.

OBSERVATIONAL STUDY

We designed a study of mobile video conferencing focused on two questions: first, what is the nature of the mobile camera work in supporting the communicative intentions of collaborators; second, what communicative problems do collaborators encounter trying to complete collaborative tasks, and how do they overcome these difficulties?

Participants

Nine pairs of adult participants (18 participants; 11 females) were recruited online and through printed ads. We recruited pairs of participants together, ensuring that one was familiar with our university campus while the other was not. With the exception of one pair (Pair 1), participants in each pair knew each other prior to participating in the study.

Method

We designed a set of four tasks to be completed by each pair. To mimic a typical mobile scenario where one person is 'out and about' and one is at home [1], we designed tasks where one participant needed to physically go to several parts of our university campus. Each task was designed so that participants needed to work together to exchange information to be successful. The desktop collaborator (DC) sat at a desktop computer with a high-resolution monitor, a webcam, a microphone, and a pair of headphones. The mobile collaborator (MC) used an Android smartphone with front and back cameras and a pair of earbuds. The MC was allowed to switch between landscape and portrait mode, and front and back cameras, as they liked. The pair was connected by video and audio using Google Hangouts.

We were interested in different behaviours given the different kinds of tasks, we limited participants to only ten minutes for each task. Even so, the entire duration of the study lasted about 90 minutes, owing in large part to the MC needing to move to different on-campus locations for each task. After completing all tasks, we conducted brief interviews with each pair to understand their experiences.

Tasks

Our tasks were based on the related work, where we aimed to keep the face validity high. These four tasks varied in three dimensions: (1) *knowledge*—in three tasks, one participant explicitly needs to share information with the other; (2) *physical movement*—in three tasks, the MC needs to walk between places; and (3) *target distance*—we anticipated that participants would use the video to share visual information, and so we varied the distance at which these targets would appear (based on each the task).

Task 1: Collaborative Physical Task. The MC constructs a MEGA BLOK structure (and has all the pieces), but only the DC has instructions on what to build. The DC's role is to guide the MC. The DC is given a set of six pictures showing the structure from different angles, but is not allowed to show these images to the MC. This task is based on the "collaborative physical tasks" commonly found in the literature (e.g., [10, 17, 21]). This task helps to provide a basis for comparison with prior work and our experiences with mobile video. Further, it mimics many of the scenarios in which a collaborator asks for help with a physical task, as seen in the related work (e.g., mechanical repair [1]).

Task 2: Campus Tour. The MC takes the DC on a brief tour of the central part of campus, showing the DC five key landmarks. The DC is to learn the spatial relationships between these landmarks. DCs were asked to sketch a map of this part of campus to illustrate their understanding of these spatial relationships. This task mimics a scenario in which one is spatially orienting or guiding someone through an environment (e.g., house tour [1]). We were interested in seeing how landmarks (and the spaces between them) would be shown and talked about in the video scene, and in particular, how MCs would move through space.

Task 3: Detail Search. The MC begins in the food court on campus, and together, the DC and MC construct a nutritious, three-day meal plan with a strict budget. The MC's role is to show the DC the different food outlets and help the DC make decisions about what would appear in the meal plan. We were interested in how participants would share different kinds of information, such as textual data (menu placards), as well as tangible objects (e.g., food items), and how this would relate to their movement through the food court.

Task 4: Negotiation – Shopping Together. The MC begins at the bookstore, and works with the DC to collect a set of gifts for a mutual friend's upcoming graduation. The team is given a strict budget, and each collaborator is given a short list of items that the friend likes or needs—each list containing items that the other list does not have. Both can share the knowledge they have, and together need to decide which gifts to buy under budgetary constraints. This task mimics help and assistance in retail shopping environments.

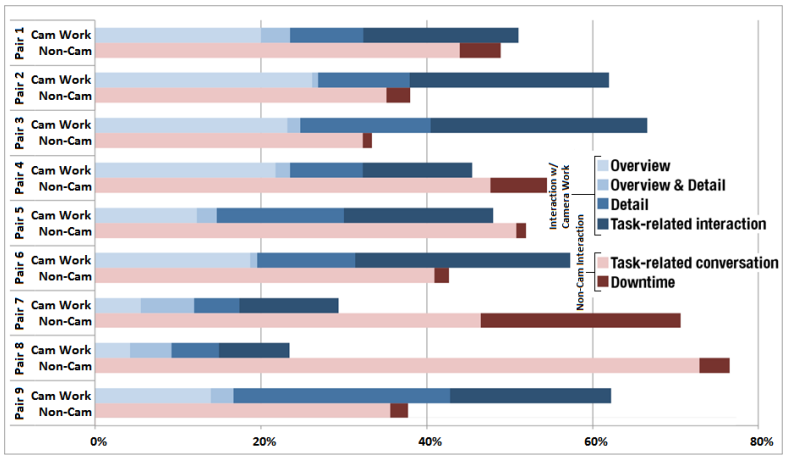


Figure 2: How participants used their time. About half of this interaction (blue-shaded) involved the camera in some way.

Data Collection and Analysis

Video and audio from both sides of the call were recorded, with an additional camera capturing the DC. We also collected field notes and videotaped interviews. These informed our analytic process.

We used *video-based interaction analysis* [13] to analyze the video data. Our analytic interests followed many of the foci outlined in [13]: specifically, the structure of events, the use of artifacts to structure events (namely, camera and objects being pointed at and referred to), how activity is organized, turn-taking, and participation structures. We applied this analysis framework by considering each of these concepts in relation to the video coding, iteratively and provisionally analyzing the data as it was collected. We reviewed this as a team, and concentrated on aspects of the framework where the mobility of the device made a difference in contrast to prior work in video-mediated communication spaces.

We observed participants' behaviours, such as how they communicated through the video connection, how the MC operated the camera, and how participants responded to the information they received. We looked for behaviours that were common across groups, as well as behaviours that were unusual. Raw videos from each session were synchronized and combined. We then iteratively annotated our videos, beginning with a small set of codes informed by interviews (e.g., showing overview/detail; using front/rear camera) and refining these codes by modifying, adding and removing codes as the process continued.

FINDINGS

We first report observations on how MCs held the phone. We then describe the basic mechanics of the MCs' camera work, in particular, how MCs try to provide overview and detail views. Then, we show how camera work supported communicative activity by supplementing speech. From here, we describe how MCs attempted to provide a sense of spatiality to DCs through movement and camera work. Finally, we discuss how the asymmetry in camera control affected participants' behaviours, and describe some of the social awkwardness experienced by our participants.

Camera Orientation. MCs held the phone in both portrait (63% of the time) and landscape (37% of the time) modes during the tasks. Landscape mode provides a wider FOV than portrait mode, and so it was surprising that MCs held the phone in portrait more often than in landscape—particularly in Task 2, where participants were explicitly trying to help orient the DC to the environment and where the larger FOV of landscape mode might have been more helpful (in this task, the split was 78%-22% in favour of portrait mode). According to participants' reports during the interviews, both social and ergonomic factors played into this decision. First, MCs found it awkward to hold the phone in landscape while walking around in public, because landscape typically signals to others that they were going to take a picture or video, whereas portrait mode attracted less attention. Second, participants found it physically awkward to hold the phone in landscape mode.

Front vs. Rear Camera. MCs used the back camera far more often than the front camera (81% of time using back camera vs. 19% front camera). Participants reported during the interviews that the DC does not usually need to see the MC's face while completing a task; instead, it was far more important to see what was in the environment. Only two groups used the front camera for more than one task; two other groups that tried using the front camera reverted to using the rear camera almost immediately.

With the exception of Task 1, MCs always held the phone in their hands; typically this was so the DCs could have a view of something (the environment, an object, or the MC—as a 'selfie'). Because MCs needed to complete a physical task in Task 1, some MCs would occasionally set the phone down or try to prop up the phone on the table.

Mechanics of Mobile Camera Work

As illustrated in Figure 2, most pairs made use of the cameras extensively to complete the tasks. We use *camera work* to describe how MCs operate the phone camera to capture the scene during the task. With the exception of Task 1 (where MCs were sitting at a table), MCs could freely manipulate the camera view—both by moving the camera phone and by physically moving. Here, we describe the basic mechanics of this camera work—first, we discuss how MCs tried to provide DCs with an overall sense of the environment (overview mechanics); second, how MCs try to give DCs an understanding of an object or landmark of focus (detail mechanics), and finally, task-related interaction, where MCs used the camera (through movement and positioning) to answer or ask questions, or to provide meaningful awareness into their action.

Camera Work: Overview Mechanics

Participants achieved overview shots in several different ways. In our analysis, we found that for these shots, either the camera was moving or held in a fixed posture, and either the MC was moving or stood still. While these practices were somewhat more fluid, the associated discussion would change depending on the overview mechanic being used.

Static overview: the MC provides an overview without moving the camera or physically moving through the space. We typically saw that MCs would show an object or several objects from far away. This occurred most frequently in Task 2, where MCs were trying to orient the DC through the space. Here, MCs would show a landmark from far away (approximately 10 to 150 metres) and keep the camera fixated while providing a verbal description of the landmark without approaching it. It seems that MCs were trying to allow DCs to "take in" the environment, as if it were a picture, allowing the DC to study features and aspects of the view. For example (Group 3, Task 2):

TIME	VERBAL	ACTION
14:28	**MC:** So, right there is the library.	*Cam:* Pointed at the library (~100 m).
14:32	**MC:** That's where you can check out books.	
14:34	**DC:** To the left of MacHall?	
14:35	**MC:** Yeah.	
14:42	**DC:** Alright. **MC:** Great!	

In Task 1, MCs would hold the phone away from the scene or step back from the desk a few feet to give an overview of the block structure at its current stage or of the available blocks. Similarly, in Task 3, MCs would point the camera at a restaurant or a picture menu board from far away to give an overview of the restaurant's offerings. In Task 4, MCs often pointed the camera at a set of items placed on a shelf and stepped back to show all of the items in frame.

Approach overview: the MC provides an overview by moving around in the environment, but holds the phone camera steady. This typically involved putting an object of focus in frame while approaching it and keeping it in frame. Usually, the MC starts doing this from far away—either because the object or landmark is simply far away, or it needs to be understood within the broader physical context—and approaches it until the viewer can see clear details of it. For example (Group 9, Task 1):

TIME	VERBAL	ACTION
12:21	**MC:** So this is Campus Security.	*Cam:* Facing entrance to the Campus Security office (~20 m).
12:26	**MC:** I'm just going to walk to it.	*MC:* Walks toward entrance; cam still has office centred.
12:30	**MC:** So if you have any security issues, you'll want to go there.	*MC:* Stops. *Cam:* centred toward on security office (~5 m).

Notice here that static and approach overviewing have different functional aims—whereas static overviewing allows the DC to take in the scene, approach overviewing is about setting an object (or its details) within context.

Camera-moving overviews: the MC provides an overview by moving or turning the camera. We identified two types of mechanics: *pan* overview and *spin* overview. *Pan* overview involves panning the phone camera across a set of objects or landmarks to give the DC a sense of what is there. This type of overview was especially common when the MC was trying to communicate a set of options to the DC. For example, MCs panned in Task 3 to show the food items a restaurant had on the counter, in Task 4 to show the items on a store shelf, and in Task 1 to show the set of available bricks. A related mechanic was the *spin* overview, where the MC provides a 180 to 360-degree overview of the environment from her current location by rotating the camera—usually while standing still. Spins were not very common (used by only three of nine MCs), but were notable due to being a means of providing a rich overview of an area. For example (Group 1, Task 3):

TIME	VERBAL	ACTION
22:25	**MC:** I'm in the food court right now.	*Cam:* Pointed at one side of the food court.
22:30	**MC:** So this is it.	*MC:* Physically spins steadily 270°. Many restaurants from both sides can be seen.

Walkthrough overview: the MC walks past a set of objects and points the camera left and right, at each one of them. For instance, in Task 4, MCs walking through clothing aisles would point the camera left and right to show the different types of clothes that were available. In Task 2, MCs also often showed DCs different landmarks as they walked past. Much like the pan overview, the purpose seems to be to provide the DC with a sense of the objects in the environment, but necessitates physically moving because objects cannot be captured all at once.

Camera Work: Detail View Mechanics

MCs also provided DCs with detail views whenever appropriate—for instance to discuss features of an item at a store, or to inspect specific food items. We saw this most frequently when there was substantial discussion of items, and a close-up view aided the interaction.

We typically described this behaviour as *centre-staging*, where a specific object or landmark is made the central focus by placing it in the centre of the frame and making it clearly visible to the viewer. For example, in Task 1, MCs centre-staged single blocks by picking them up and holding them in the centre of the frame (e.g., to ask whether the block was the correct one). In Task 2, MCs centre-staged landmarks while describing them. Similarly, in Tasks 3 and 4, MCs centre-staged restaurant signs, menu boards, store and food items, and price tags—for instance, one MC held the camera close to a package of sushi to give her partner a view of its contents. Another example (Group 9, Task 4):

TIME	VERBAL	ACTION
36:45	**MC:** It's called "Aquabee Sketchbook."	
36:53	**MC:** Here, this is the name.	*Cam:* Pointed at the label on the front of the sketchbook.
37:00	**DC:** Okay.	*Cam:* Moves away.

In some cases, the MC could pick up an object in question (e.g., in Tasks 1 and 4), and move it to the centre of the frame. In Task 1, six of nine MCs picked up the assembled set of bricks and rotated it to allow their partners to see it from different perspectives. This was done to both provide a view of the "current" state of the structure and to provide assistance in figuring out what steps to take next.

Camera Work: Technology Limits

Overview mechanics each had different purposes, but many involved trying to share something about the environment (e.g., where landmarks were relative to each other, the space of possibilities, etc.). MCs overcame the poor FOV of the camera (further exacerbated by MCs holding the camera in portrait mode) through the use of overviews constructed through movement (e.g., pan, spin, approach)—yet, there are obvious problems with these. In particular, they require the DC to integrate multiple scenes over time, remembering what they saw in relation to what they are seeing now. Furthermore, the video had substantial motion blur, as the camera could not focus quickly enough. Based on what we saw, it was clear that while the MCs were well intentioned in constructing overview shots, they were not always effective. For instance, DCs rarely got a complete sense of the entire spatial context. This was evident in the maps drawn by DCs in Task 2, which bore little resemblance to areas themselves. Similarly, with the walkthrough approach, it was easy to forget about objects that were no longer in view.

Because of the low resolution of the video feed from the mobile phone (and the lack of zoom controls and manual focusing), MCs also often had a hard time showing details from distances that would be appropriate if the DC were collocated. Instead, to accommodate the poor view of the camera feed, MCs needed to walk closer to objects of interest. For example, in Task 3, MCs would sometimes read restaurant menu boards aloud in addition to pointing the camera at them. They did this because they knew that it was hard for their partners to see the details of the menus or even read from them. In principle, with a high enough resolution, or the ability to bring attention to the area of interest, these extra movements would not be necessary.

Camera Work and Communication

While camera work supplemented speech, occasionally it also seemed to replace conversation altogether. Specifically, we saw all MCs ask and answer some questions solely through their manipulations of the camera view. For example, in Task 1, the MC asks where to place a block by hovering it over several possible spots (or in different orientations)—moving through options quickly enough that it is clear the MC is waiting for the DC to simply interrupt and say "that is where it should go," rather than waiting for a response to each possibility. Similarly in Task 3, the MC hovers the camera over several food items as a way of asking if the DC wants to try any of them. Again, the MC waits for the DC to interrupt the movement as a way of acknowledging positively which food item should be selected.

Similarly, MCs answered questions through camera work—responding, for instance, to questions from the DC by simply pointing the camera at an object or allowing the movement of the camera to indicate that an action is being taken. In Tasks 3 and 4, for instance, when DCs asked for the price or physical description of an item, some MCs would simply point the camera at the item or its price tag. For instance (Group 3, Task 3):

TIME	VERBAL	ACTION
19:28	**MC:** You want a muffin? It costs $1.43.	*Cam:* Pointed at the menu board (~5-7 m).
19:31	**DC:** Really? How big is it?	
19:33		*MC:* Slowly approaches the muffin on the display counter.
19:38	**MC:** This big. **DC:** Oh, okay!	*Cam:* Focused on the muffin on the display counter (~few cm).

To direct the DC's attention, MCs would sometimes point the camera toward an object of interest; however, this was sometimes inadequate—particularly if many things could be seen or the video scene was cluttered. Consequently, MCs would sometimes point through the frame (i.e., point their finger in front of the camera) to direct attention to certain things (e.g., Figure 3). While this was effective, it was very cumbersome for the MCs.

Conveying Spatiality through Movement in Space

MCs seemed to be aware that DCs would not be able to understand the spatial layout of the environment with a simple walkthrough. Consequently, many MCs adopted awkward movement and conversation strategies to help DCs understand the environment. For example, some MCs tried not to turn the camera around while moving (e.g., by pointing only in the direction that they were moving), while others would turn slowly while reporting loudly what they were doing. Other MCs moved only in a grid-like fashion (i.e., in simple directions such as left, right, forward, and backward, rather than diagonally), to make it easier for their partners to understand their movements. Figure 4 contains a sketch depicting an observed example of these actions.

MCs would also show new landmarks in relation to other familiar landmarks. There were two strategies that MCs used to accomplish this: *anchoring* and *backtracking*. With *anchoring*, the MC refers to a previously seen (or previously visited) landmark. This was done is a few ways:

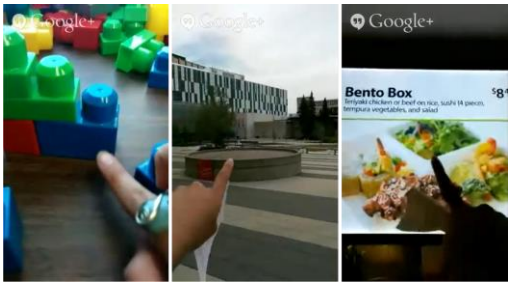
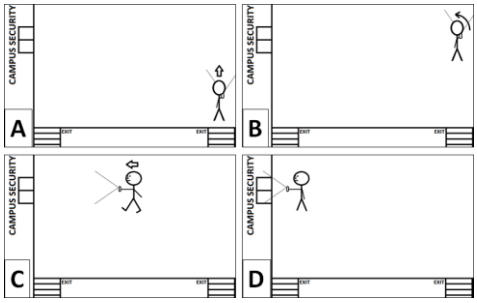


Figure 3: Pointing with hands through the video frame.



	TIME	VERBAL
A	11:15	**MC:** Okay, so I'm walking in [to the building].
B	11:23	**MC:** And then, I'm turning left.
C-D	11:30	**MC:** And this is what [the security office] looks like.

Figure 4: MCs often moved slowly and in a grid-like fashion; while pointing the camera forward and carefully describing their movements.

by pointing the camera at the landmark, by turning the camera at and pointing at the landmark, and by simply saying where the landmark was in relation to the MC. This technique was used at least once by every MC. Similarly, *backtracking* refers to when an MC goes to another location by walking backwards in a path that she took before. By doing this, the MC could recap the locations of previously seen landmarks while approaching a new landmark. Some MCs chose to backtrack rather than take a new path, even when a new path was available and more convenient, because they thought that it would be easier for their partners to understand. Both these strategies rely on camera work—on the MC's part to point and show physical relationships between landmarks, and again relying on the DC to be able to develop a mental model of the space based on the video.

In spite of all of these strategies, DCs generally developed poor (and incorrect) understandings of the spatial environment.

Asymmetries of Control, Participation and Awareness

The MC had complete control over the camera's location and orientation, meaning that the DC's ability to perceive the environment was strictly mediated by the MC. This had some interesting consequences for how the DC and MC interacted with one another—specifically when the DC wanted to direct either the action or the camera view.

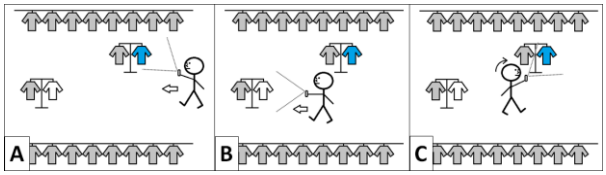
Referring to Objects. In Task 1, only the DC knows how the final structure should look—as such, they explicitly need to provide information and instruct the MC on how to complete the task. In practice, this frequently amounted to describing a block and where to place it (along with the orientation). When participants are collocated, these instructional tasks tend to be fairly straightforward (e.g., [10, 17]). In our study, the explicit lack of access to the MC's workspace presented some serious problems for DCs.

DCs gestured at the computer screen (e.g., in Task 1 to refer to a specific block, or in other tasks to direct the action within the mobile frame). Unfortunately, as illustrated in

Figure 5: Right: The DC's gestures at the screen cannot be seen by the MC. Top left: The DC's view. Bottom left: The MC's view of the DC.

Figure 5, these gestures were not visible to the MC, meaning their communicative intent was lost. Nevertheless, these gestures were quite frequent—seven out of nine DCs made such gestures and, of these, averaged ten times during Task 1. The inability to view gestures (which included pointing, swiping, and rotating gestures) meant that the rich communicative intent was lost. DCs in Task 1 were frequently frustrated by their inability to articulate what is very easily communicated through gesture.

Negotiated Camera Control. DCs also had no control over the camera view—as such, if a DC wanted to see something in the view, they needed to negotiate with or ask the MC. There were a few instances when the DC interjected verbally and asked the MC to focus on or point the camera toward something specific in the environment. Some MCs seemed to provide opportunities for such interjections when panning slowly over a set of items or around the environment, sometimes asking questions while performing the camera movements. This was rarely smooth though. For example, in the vignette illustrated by Figure 6, while the MC is doing a walkthrough, the timing of the DC's requests are sufficiently behind with what is actually in the mobile frame that the DC's requests rarely get fulfilled. Here, while the pair is deciding on a shirt to buy, the DC says "the blue one"; however, at this point the blue shirt is out of frame and a white shirt is in frame. As the MC is turning around to point back at the blue shirt, the DC says "yeah, the white one"; but at that point the white shirt is out of frame and the blue shirt is in frame again. Note that the verbal references typically were in broad terms about what



	TIME	VERBAL
A	12:49	
B	12:52	**DC:** The blue one!
C	12:53	**DC:** Yeah, maybe the white one! **MC:** This one? *(Referring to blue shirt.)*

Figure 6: MCs' attempts to allow the DC to interject do not always turn out smooth.

was seen (e.g., colour, shape, type of object, etc.). In three cases, the MC ignored these interjections or did not understand what the DC was trying to direct focus to.

Such interjections did not occur frequently—only occurring once or twice for six of the nine pairs. One reason could be that the DC gets a very limited view into the MC's environment and only interjects if something interesting shows up in frame. Another possibility (as illustrated in the vignette from Figure 6) is that ultimately such interactions are frustrating for both the DC and MC, and so participants may have been trying to minimize conflict.

MC Takes Over Completely. Because of the limited opportunities for the DC to interject or guide the MC's movement or focus, three MCs ended up completely taking the lead in one or more of Tasks 3 and 4 (Pairs 4, 7, and 8)—tasks designed for discussion and working together. For instance, one MC simply decided unilaterally what the DC was to order and eat from the food court in Task 3. Similarly, four MCs simply took on most of the shopping of Task 4 by themselves, going wherever they wanted and looking through items without showing or even describing them to their partners. In these cases, the DC ended up providing information only verbally or only discussing issues brought up by the MC.

Limited Awareness. The MC rarely paid close attention to the DC's view, and as such pairs would often miss opportunities to coordinate. In several instances, the MC tried to show something or do something expressly for the DC through camera work, but the DC did not notice. For example, in Task 3, while a pair was discussing healthy eating options, the MC slowly panned across a set of doughnuts (likely as a joke). Unfortunately, the DC was looking at her paper notes and did not notice.

Social Awkwardness and Distraction

Our MCs described feeling awkward holding the mobile phone in public while in the video call. In particular, many felt that holding the phone in landscape drew substantial unwanted attention as it implied or signaled the taking of a photo or video. As a consequence, MCs often held the phone in awkward ways (Figure 7): either close to one's body and away from the face (Figure 7 left and centre), making it difficult to see the screen; or close to one's chin (Figure 7 right), with the back camera pointed down at the floor and the front camera pointed up at the MC's chin.

An interesting consequence of the phone-holding postures was that MCs became very engrossed in providing their partners with a "good view." This meant that MCs frequently failed to notice other people or hazards around them. One participant was nearly hit by a vehicle passing by. Such incidents also suggest that, for at least some tasks, a wearable camera may be a good solution.

Figure 7: MCs held the phone in awkward ways.

DISCUSSION

In articulating the mechanics of mobile camera work, our intention was to tease apart the goals of actions from the constraints placed on the MC by the design of the technology. Table 1 summarizes the observed goals participants had while completing the tasks, and the specific actions they would take to achieve these goals. Along with each goal, we present the informational challenges of the mobile and remote parties. This table acts as a tool in the discussion of mobile video collaboration practices and needs. Next, we discuss several implications of these findings. For simplicity, we continue to use DC and MC; however, more generally, DC refers to the party not controlling the camera view in question.

Many of the findings we outline here confirm prior findings from Gaver et al. [9]; however, the fact that a person is in explicit control of a mobile camera here adds substantial nuance. In particular, we see unpacking a user's *goal* with a particular camera angle (e.g., overview, detail, etc.) as adding something interesting to the overall discussion. In particular, if we can establish what people are trying to do, there may be opportunities for designing certain kinds of automation that can help address these goals.

Gesturing and Drawing Attention. Collaborators wanted to show details of the scene and reference objects to their partners during all of the tasks. We saw awkward attempts to gesture at the video scene (e.g., MCs putting their hands in the video scene), and the DC's gestures failing entirely (because their gestures could not be seen). Designs should provide effective and appropriate means to gesture. While this idea is not new (e.g., [6, 7, 8, 11, 16, 17]), most mobile video conferencing systems still do not provide this support. Further, even when gestures are observable, they might be missed, partly due to reduced awareness (e.g., because of the small mobile phone screen). In addition, while previous work has explored gesturing mechanisms that rely on explicit messaging (e.g., [6, 7, 8, 11, 17]), our findings suggest that it may also be important to support gesturing through social cues (e.g., by pointing, as in [16]). Furthermore, our findings also suggest that it is important for MCs to be able to read these cues while still maintaining high awareness of and concentration on the surrounding environment. This may be possible, for example, through vibrotactile wearables, where the vibrotactile device can be used to indicate which direction someone is interested in looking.

Goal	Camera Work Mechanic
Show an environment/scene *Challenge:* DC should be able to explore the scene him/herself	Static overview, spin overview
Show a set of alternatives (objects) *Challenge:* DC should be able to remember all alternatives	Pan overview, walkthrough
Show detailed information *Challenge:* MC should be able to do this quickly and efficiently	Centre-staging, walking close to object, picking up object
Referring to an object *Challenge:* MC should be able to do this quickly and efficiently *Challenge:* DC may not be able to see the object	Pointing through the video scene, verbal interjection
Provide spatial awareness *Challenge:* DC cannot explore the scene him/herself	Backtracking, anchoring

Table 1: An MC's goal and the associated camera work observed.

Visibility of Objects. It was not always easy for either party to reference objects or scenes, either as a communicative act or to aid in memory. This was particularly true when the MC was very close to or far from a target object, or when a scene was not easily framed. This necessitated careful camera work to get close enough to an object to centre-stage it or to get in a position to correctly frame a scene. Two things make this procedure challenging: first, ensuring that the object of interest is in the scene (so the other person can see it); second, bringing the other collaborator's attention to the object, or the specific part of the object as necessary. Because the MC has control of the camera view, it is easier for them to accomplish this. Designers need good ways to provide DCs a means to convey what camera shot is needed. Allowing DCs to review recent scenes might save them from having to remember details or to request the MC to reframe a scene. This idea has been touched on to some extent by previous work (e.g., [7, 8, 16]).

Spatiality and Context. We also saw that both MCs and DCs are interested in understanding spatial context—to contextualize gestures and comments, and to understand what else is in the MC's environment that is not conveyed in the video scene. For DCs, because they did not have this context, they were unable to make coherent suggestions about what camera shot was necessary, or meaningfully help with decision-making (e.g., in the shopping scenario); instead, they were strictly confined to what the MC happened to be pointing the camera at. To address this, a larger FOV camera would likely again help (much like the workspace-oriented camera of [3]). This could provide the DC with better awareness of the MC's environment, allowing for more opportunities for the DC to interject and guide the MC's movement or focus. While our study did not lead to any concrete evidence that a limited FOV hinders collaboration, some participants did mention that they felt the limited FOV hindered their ability to complete certain tasks. We leave further investigation of this for future work.

Remote Control. We observed that MCs recognized that building a spatial awareness would be challenging for the DCs, and adopted a number of ways to accommodate (e.g., backtracking and anchoring). However, these failed to allow the DCs to build an accurate mental map of physical spaces. If the DCs had been local, this could have been easily achieved by looking around a space. However, this was not always possible because a particular view was not available when a DC needed it, and requesting it from the MC was awkward and cumbersome. We also saw that the lack of control for DCs led to frustration. Previous work has proposed and explored means to allow DCs to control their view of a mobile scene both directly (e.g., by controlling a camera [30] or a view of a reconstructed representation of the scene [7, 8, 16]) and indirectly (e.g., by providing means for DCs to request what direction to point the mobile camera in [16]). Our findings suggest that the social signals we use in day-to-day interaction should be respected while providing these means. In other words, approaches that rely on implicit and natural social cues such as pointing (e.g., [16]) are more favourable than approaches that rely on explicit user intervention (e.g., [7, 8]).

Avoiding Social Awkwardness. We observed participants holding their phones in awkward ways and avoiding the use of landscape mode because it communicated to others that they were taking photos or videos. Designers should consider ways to communicate to observers that a person is engaged in a video conversation, allowing others to understand what an MC is doing (e.g., by flashing the camera light at regular intervals). This might also communicate to people in the environment that this person's attention is divided and may allow others to accommodate (perhaps avoiding collisions).

Wearables. Collaborators had a difficult time manipulating the phone camera while completing Task 1, which required handling other physical objects. Hands-free mobile video conferencing technologies should be designed to support collaboration in tasks that require both hands. These types of technologies have been explored by other researchers (e.g., [22, 24, 30]). However, these technologies present interesting trade-offs in relation to the problems described above—as they are likely only able to provide detail views rather than good contextual overviews. Furthermore, being coupled strictly to a body part, be it a chest (as a pendant), or one's head (as a heads-up view) means that the FOV is strictly limited to what the MC is looking at, and we have observed that this can be frustrating for the person not holding the camera.

CONCLUSION

Current designs of mobile video conferencing technologies are mainly the mobile equivalent of their desktop counterparts. Very little work has gone into understanding exactly what changes in mobile video conferencing scenarios, particularly in support of shared tasks. Yet, mobile video enables a whole host of new applications and scenarios that were previously unavailable (e.g., aiding in mechanical repair, supporting navigation, etc.). As we have seen, it is likely that these scenarios are poorly supported by current tools. We saw that participants frequently found it difficult to effectively complete tasks—in part because of

communicative breakdowns in relation to the camera view. Poor FOV and asymmetry of control mean that people cannot equally contribute to ongoing interaction.

Our study of mobile video conferencing has provided new insights on the ways in which camera views are used to help support communication across a video link. We have provided the first articulation of the mechanics involved in completing collaborative tasks using current mobile conferencing systems. Based on this new framework built from observation, we have outlined key challenges and several implications that could help designers build and improve mobile video conferencing tools in the future.

ACKNOWLEDGEMENTS

We thank all of our study participants, as well as Ray Sharma of XMG Studio, NSERC, SurfNet, GRAND NCE, Nokia, and AITF for their generous support of this work.

REFERENCES

1. Brubacker, J. et al. (2012). Focusing on Shared Experiences Moving Beyond the Camera in Video Communication. In *Proc. DIS 2012*, 96-105.
2. Clark, H.H., and Marshall, C.R. (1981). Definite reference and mutual knowledge. In *Elements of Discourse Understanding*, 10-63.
3. Fussell, S.R. et al. (2003). Effects of Head-Mounted and Scene-Oriented Video Systems on Remote Collaboration on Physical Tasks. In *Proc. CHI 2003*, 513-520.
4. Fussell, S.R. et al. (2003). Where do Helpers Look?: Gaze Targets During Collaborative Physical Tasks. In *EA CHI 2003*, 768-769.
5. Fussell, S.R. et al. (2003). Assessing the Value of a Cursor Pointing Device for Remote Collaboration on Physical Tasks. In *EA CHI 2003*, 788-789.
6. Fussell, S.R. et al. (2004). Gestures Over Video Streams to Support Remote Collaboration on Physical Tasks. *HCI 19(3)*, 273-309.
7. Gauglitz, S. et al. (2012). Integrating the Physical Environment into Mobile Remote Collaboration. In *Proc. MobileHCI 2012*, 241-250.
8. Gauglitz, S. et al. (2014). World-Stabilized Annotations and Virtual Scene Navigation for Remote Collaboration. In *Proc. UIST 2014*, 449-459.
9. Gaver, W.W. et al. (1993). One is not enough: multiple views in a media space. In *Proc. CHI 1993*, 335-341.
10. Gergle, D. et al. (2004). Action as Language in a Shared Visual Space. In *Proc. CSCW 2004*, 487-496.
11. Gutwin, C., and Penner, R. (2002). Improving Interpretation of Remote Gestures with Telepointer Traces. In *Proc. CSCW 2002*, 49-57.
12. Inkpen, K. et al. (2013). Experiences2Go: Sharing Kids'Activities Outside the Home with Remote Family Members. In *Proc. CSCW 2013*, 1329-1340.
13. Jordan, B., and Henderson, A. (1995). Interaction Analysis: Foundations and Practice. *Journal of the Learning Sciences 4* (1), 39-103.
14. Judge, T.K., and Neustaedter, C. (2010). Sharing Conversation and Sharing Life: Video Conferencing in the Home. In *Proc. CHI 2010*, 655-658.
15. Judge, T.K. et al. (2011). Family Portals: Connecting Families Through a Multifamily Media Space. In *Proc. CHI 2011*, 1205-1214.
16. Kasahara, S., and Rekimoto, J. (2014). JackIn: Integrating First-Person View with Out-of-Body Vision Generation for Human-Human Augmentation. In *Proc. AH 2014*.
17. Kirk, D., and Fraser, D.S. (2006). Comparing Remote Gesture Technologies for Supporting Collaborative Physical Tasks. In *Proc. CHI 2006*, 1191-1200.
18. Kirk, D. S. et al. (2005). Ways of the Hands. In *Proc. ECSCW 2005*, 1-21.
19. Kirk, D. et al. (2007). Turn It This Way: Grounding Collaborative Action with Remote Gestures. In *Proc. CHI 2007*, 1039-1048.
20. Kirk, D. et al. (2010). Home Video Communication: Mediating 'Closeness'. In *Proc. CSCW 2010*, 135-144.
21. Kraut, R.E. et al. (2003). Visual Information as a Conversational Resource in Collaborative Physical Tasks. *HCI 18* (1), 13-49.
22. Kuzuoka, H. (1992). Spatial workspace collaboration: a SharedView video support system for remote collaboration capability. In *Proc. CHI 1992*, 533-540.
23. Licoppe, C., and Morel, J. (2009). The Collaborative Work of Producing Meaningful Shots in Mobile Video Telephony. In *Proc. MobileHCI 2009*, 254-263.
24. Mann, S. (2000). Telepointer: Hands-free completely self-contained wearable visual augmented reality without headwear and without any infrastructural reliance. In *Proc. ISWC 2000*, 177-178.
25. Norris, J. et al. (2012). CamBlend: An Object Focused Collaboration Tool. In *Proc. CHI 2012*, 627-636.
26. O'Hara, K. et al. (2006). Everyday Practices with Mobile Video Telephony. In *Proc. CHI 2006*, 871-880.
27. Ou, J. et al. (2003). DOVE: Drawing over Video Environment. In *Proc. MULTIMEDIA 2003*, 100-101.
28. Procyk, J. et al. (2014). Exploring video streaming in public settings: shared geocaching over distance using mobile video chat. In *Proc. CHI 2014*, 2163-2172.
29. Raffle, H. et al. (2010). Family Story Play: Reading with Young Children (and Elmo) Over a Distance. In *Proc. CHI 2010*, 1583-1592.
30. Sakata, N. et al. (2003). WACL: Supporting Telecommunications Using Wearable Active Camera with Laser Pointer. In *Proc. ISWC 2003*, 53-56.
31. Sodhi, R.S. et al. (2013). BeThere: 3D Mobile Collaboration with Spatial Input. In *Proc. CHI 2013*, 179-188.
32. Tang, J.C. et al. (2013). HomeProxy: Exploring a Physical Proxy for Video Communication in the Home. In *Proc. CHI 2013*, 1339-1342.
33. Virtual Photo Walks. http://www.virtualphotowalks.org.
34. Yarosh, S. et al. (2010). Video playdate: toward free play across distance. In *Proc. CHI 2010*, 1251-1260.