CamBlend: An Object Focused Collaboration Tool

James Norris, Holger Schnädelbach
Mixed Reality Laboratory
University of Nottingham,
Nottingham, NG8 1BB
{psxjn, holger.schnadelbach}@nottingham.ac.uk

Guoping Qiu
Intelligent Modeling and Analysis
University of Nottingham,
Nottingham, NG8 1BB
guoping.qiu@nottingham.ac.uk

ABSTRACT
CamBlend is a new focus-in-context panoramic video collaboration system designed to facilitate the interaction with and around objects in a lightweight, flexible package. As well as the ability to view very high resolution local and remote video that covers a full 180° field of view, the system contains a number of tools which facilitate bi-directional pointing between two remote spaces. In the first quasi-naturalistic exploratory study on a focus-in-context video system, we show a number of unique object referencing behaviours, including un-intentional or 'implicit' pointing and a number of scenarios where this was advantageous. Additionally the study highlighted some of the problems inherent in aligning between screen-based and real-world perspectives.

Author Keywords
CSCW; collaboration; interaction analysis; focus+context

ACM Classification Keywords
H.4.3. Communications Applications: Computer conferencing, teleconferencing, and videoconferencing

INTRODUCTION
Video communication technologies are widely used by organisations to support teams in remote collaboration [14], whether this is on the desktop or in the boardroom [22]. Whilst such tools can and are frequently used for purely discursive ends it is quite common to wish to use video-technologies to support a collaborative task at a distance which might entail access to the connected/shared spaces, specific artefacts and resources located in those spaces and of course the team members involved. In the majority of video communication set-ups in use today, this access is impeded by principal technical limitations.

Spatial inconsistencies are introduced because of the anisotropy of video communication (different set-ups at each node create interactionally different views) and the way the pinhole camera model does not preserve lines of sight [8]. Secondly, standard cameras have a much more limited field of view (FOV) compared to human vision. These technical limitations have an impact on interaction, and the interactional issues that emerge around the shared access to distributed resources and the conduct of team members are well documented in the literature [20, 21].

CamBlend is a new focus-in-context panoramic video collaboration system that solves several technical limitations inherent in current systems that increase FOV. The system design is exploited to provide facilities which help with some of the problems generated through the spatial inconsistencies of video communication. There are two contributions, firstly in this technical system, and secondly by the first exploratory study of a focus-in-context video collaboration system.

RELATED WORK
The problems with spatial consistency are usually manifested in the inability to point to, or reference artefacts in either local (in order to show a remote participant something) or remote environments. Collaborative tasks are frequently grounded through referential statements and gestures made in relation to objects of common interest. Direct manipulation of the objects themselves is also a means by which action is coordinated and made accountable to others. This is particularly relevant for expert helper tasks [3, 6, 23] as well as more generally in workplace environments [12, 20]. These problems stem from the fracturing of interaction spaces [20] and the technical limitations in maintaining consistent spatial relationships. In this context, Fussell et al provide a succinct classification of classes of visual information necessary to ground conversations [6]. Views of participants' heads and their faces, of their bodies and actions, of shared task objects and of the work context, support joint attention, the monitoring of comprehension and conversational efficiency to varying degrees, by establishing a shared visual interaction space.

Some high-end (expensive, custom built) systems address those issues by careful architecture of the interaction environment, including HP Halo [19] and t-Room [22], but as well as being costly these systems are inflexible in how and where they're deployed. The remote-controlled GestureMan [18] system is more flexible, allowing remote pointing with the use of a laser pen, but end-users found it difficult to relate the robot to activities in the remote environment. More lightweight is a cursor pointing system [5], which although limited in its expressive capacity,

CHI'12, May 5–10, 2012, Austin, Texas, USA.

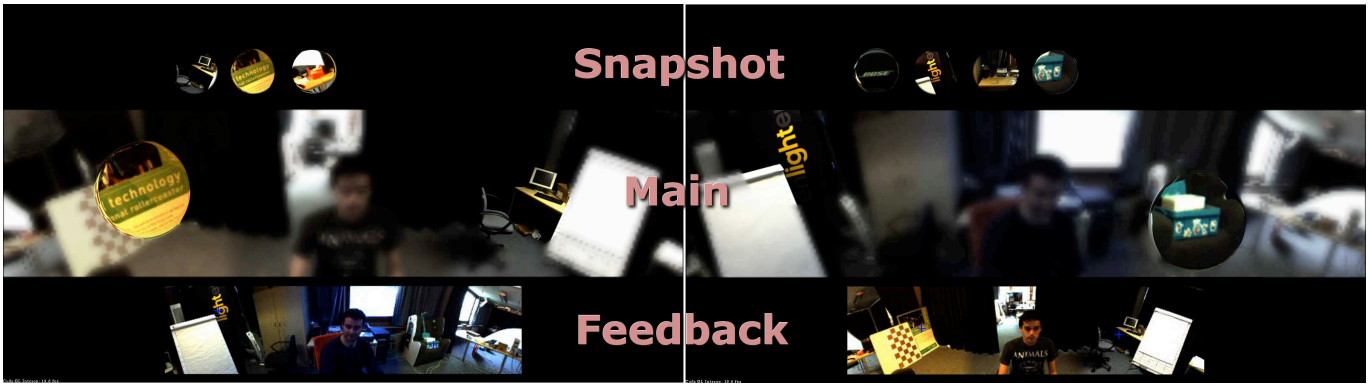


Figure 1. Two connected CamBlend nodes with snapshot mechanism (top), main remote feed and feedback (bottom)

provides a flexible solution, and the study of which showed that with it participants were faster in locating objects.

The inability to reference objects is compounded by the limited FOV of standard cameras, limiting the contextual information that end-users have. They are often unaware of objects that surround the narrow camera FOV and are therefore unable to reference them as they would in a face-to-face scenario, which has been shown in generic collaboration tasks [10, 15], expert helper systems [6] and collaboration over CVEs [13]. There are various methods to increase FOV, including fish-eye lenses, curved mirrors and camera arrays. The relationship between resolution and FOV means that they all require much more bandwidth in order to represent the fidelity of standard video. Simply using very high resolution imagery is unmanageable in three areas: in terms of live communication, as it cannot be streamed in real time; in terms of computational power required, as it cannot be encoded/decoded or otherwise processed in real time; in a display sense, as monitors cannot render the resolutions that would be required. Further, we hypothesise that systems which only increase the FOV of cameras will produce video which contains large areas of redundancy, areas that do not change and are not central to the interaction, as interactions usually occur in narrow, focused areas, unpredictably located.

Previously, this has been addressed by providing multiple but related views into the same space, in order to provide both detailed focused views and wide-angle contextual overviews [3, 9]. This work can most usefully be categorised in terms of the focus plus context literature [2], which looks at efficient ways to find areas of detail within a large dataset. Video systems in this space include for example the Polycom CX5000 (formerly Microsoft Roundtable) [4], which uses a camera array to provide a full 360° FOV for connecting meeting rooms. The system automatically selects regions of interest using directional microphones. Other systems include the Extra-Eyes [24] system, using motorised pan-tilt-zoom cameras for detail together with a wide-angle fish eye camera for overview. This design has more recently been used in a range of expert/helper experiments [3, 23], which concentrate on discovering the most effective visual information configuration and focal point control schema.

As well as problems with scalability, speed and control, a theme that runs through all the above work is that of reconciliation between different camera views. Although most pronounced in the MTV studies [9], the Extra-Eyes [24] work also reported on the performance benefits of linking detailed views and overviews. The focus plus context work by Baudisch et al [2] shows that for dynamically changing information, keeping the focused view in context allows for more effective use of peripheral information. Arguably, the closer and more 'linked' these views are, the less effort might be required in relating them.

CAMBLEND DESIGN

CamBlend is addressing the established challenge of providing flexible, cost-effective access to remote spaces, artefacts and people, by combining wide camera FOV with multiple controllable, in-context views of detail. Camblend is a video collaboration system covering 180° FOV, which represents the area of focus contextually within its background. It also contains a set of tools which provide bi-directional pointing facilities. CamBlend scales to a number of focus areas that can each be manipulated independently, and without observable latency. CamBlend is designed to be a portable video conferencing technology that could be deployed on a desktop or large screen media space, requiring a self-contained camera unit and custom software. The following describes its two main development iterations.

The Core System – ITERATION 1

At its core CamBlend is a focus plus context video system that represents areas of attention in high resolution, contextually inside a wide-angle, low-resolution background. The system is scalable in that multiple users can each own unique views into the same space.

Implemented in C++ and CUDA [1], the system creates a panoramic video stream from 3 high resolution (1024x768) cameras aligned to cover a 180° FOV. A very high resolution panorama is generated (3072x768) using the FlyCam method [7], implemented on the GPU. This high resolution panoramic image forms the base dataset, from which many concurrent focus areas can be extracted and transmitted in real-time. As shown in Figure 1, the background rendered is a very low resolution (256x60)

version of the panoramic image. Overlaid onto this, in context, are high resolution (368x368) focus windows. Users may manipulate the position of these focus windows without observable latency by dragging them over the contextual view.

It is also possible to have the focus windows zoom into the remote space, shown in Figure 1. This is performed in-context, so that areas around the focus window are not occluded. A focus plus context technique is used [17], where linear and non-linear magnification is combined in order to preserve straight lines while zoomed, but smoothly blend between the focused and contextual areas. Because the full resolution of the panorama is so large, this zoom functionality becomes a 'real' zoom, revealing extra underlying information.

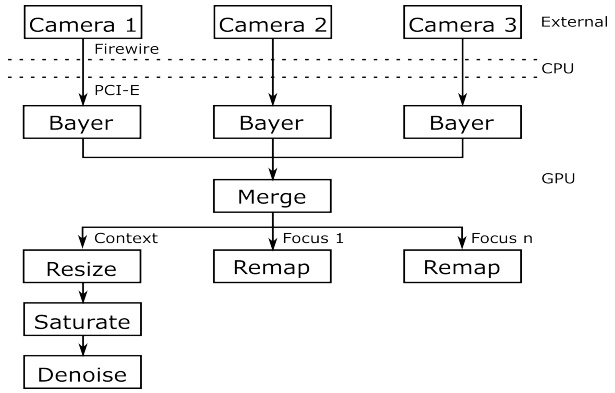


Figure 2. The video pipeline shows the processing of image data from camera hardware (top) to screen (bottom). Further processing for over the air communication (OTA) is missing. The implementation has not been necessary for the study but newer versions of CamBlend include a full OTA stream.

Until recently this design (visualized in figure 2) might only have been possible by building some custom digital signal processing (DSP) hardware. Now with the use of CUDA [1] as a GPGPU programming language, it will run on most desktop PCs. This system design improves on the PTZ camera pair design [24] in three ways; firstly movement is instant without the physical limitations of motorized cameras; secondly the focused and contextual views are seamlessly blended by being generated from the same image; thirdly there is a scalability dimension which allows the same person to have multiple views into a single space (explored in the study) and multiple people unique views into a single space (explored in principle), both of which would require multiple PTZ cameras.

Focus Groups Feedback

Iteration 1 was deployed as a single node demonstration, exposing the interaction paradigm of user-controllable focus windows, visually it matched Figure 1, but did not contain the snapshot and feedback features shown. This version was discussed with two focus groups. The aim was to collect feedback very early on in the development. Participants within both focus groups had no previous knowledge of the system or input into the development. Both sessions were audio recorded and the main points have been drawn out of this recording below.

One focus group was to a group of seven teachers at a technology-focused secondary school in the UK. The teachers were asked about their thoughts on the interaction paradigm, as well as potential use-cases within the school they could imagine. CamBlend was setup and trained to view a doorway area where people were frequently entering and exiting. The teachers were able to list a large number of scenarios where the system could be used. These included observing a remote, inaccessible location, for example an archaeological dig site from the classroom, and a tool to record, annotate and playback dance classes, where children are recorded as they perform and the clip is collectively scrutinised, annotated and saved. There were also ideas around video editing and directing, recording and later filtering to show only relevant clips, as well as suggestions to use CamBlend to connect remote classrooms for creative activities. Focus windows could be used to make groups of pupils aware of attention that they are currently receiving.

Another session was held with five researchers within our department, chosen for their HCI and development experience. As well as feedback on the interaction paradigm, this session was also to get expert advice into the design decisions required to enable the effective interaction with objects. For this session we were able to prepare the study space with some objects to interact with, including some Lego, a laptop, a white-board and some posters with various sized text. Participants were asked questions about the system and encouraged to interact with it over the course of a two-hour session. The feedback again was positive; participants had no trouble grasping the concept or interaction technique. Researchers here identified the potential of the focus windows of representing areas of attention, and suggested a feedback mechanism to allow them to understand what others see of their local space.

The focus groups highlighted the wide range of possible application areas but also drew attention to potential areas for further development. In particular, features for recording, features for making others aware of one's attention and features for view feedback were those that seemed most promising to improve support for interaction. These areas were addressed in the final version. The aim was to both respond to the very initial focus group feedback and to devise a prototype that could be studied in a laboratory setting in a formative way.

Refining Interaction Support – ITERATION 2 (Studied)

The panorama generated with 3 cameras (aspect ratio 4:1) forms the main window showing the remote scene (see Figure 1 showing two symmetrical networked nodes). The panorama is centrally placed on screen. Focus windows are floating over this contextual background and can be dragged to any position using a mouse. There is then space

above and below the panorama for additional features described briefly below.

Recording Snapshots

One advantage of segmenting the panorama for the generation of focus areas is that those streams can be easily digitised for further processing. This creates possibilities in saving, sharing or documenting video or image resources. One feature that touches on these affordances is the snapshot window. When a user clicks the right mouse button over a focus window, a screenshot is taken of the focus window area, which displays at the top of the screen in one of several slots (we have allowed for six here because of the task, but it is flexible). This feature allows users to document objects or events, which interest them. Figure 1 shows several recorded snapshots.

Feedback and bi-directional referencing

The feedback window renders at the bottom of the screen as a miniature panorama, showing the user's own space with the other user's focus window positions augmented over the top with a blue cross. The blue crosses can in turn be directly manipulated, allowing users to point in their own space. Taken together, this provides both control and awareness in both directions and provides for a shared visual interaction space, which is widely reported as being beneficial throughout the video collaboration literature [6, 9, 19]. For example, a user has the ability to look into a remote space to explore. Due to the remote users live feedback window they are also implicitly stating, "I am looking at this". This then provides the ability to implicitly point, to direct attention 'without' intention. In addition a user may also manipulate a remote user's focus window by dragging the blue cross, to state "I want you to look at this", while this can also be used to explore one's own space.

This implicit pointing ability depends on the use of the focus windows to view objects in detail. If the background provided enough detail users needn't use the windows. Conversely the background must provide enough detail to be useful as a source of contextual information, i.e. by representing motion and basic image outlines clearly. Through a series of informal iterations, a balance between these two was reached. The final system uses both a k-nearest-neighbour de-noise algorithm, and a colour de-saturation algorithm which results in a near grey-scale background image with a 'glazed' effect. A flexible network layer provides a high level TCP socket interface supporting all other control and synchronization functionality.

STUDY

A lab based formative user study was held with this version of the system. The objective was firstly to evaluate the basic interaction design and identify any interactional problems. We were particularly interested in the ways that the multiple views would be reconciled by users and the way that the focus windows would be used for awareness and pointing. In order to spur the intensive use of the system the study was designed around a lab based collaborative object finding task. Despite the somewhat artificial nature of the task, we feel it remains representative of a workplace scenario that involves the referencing of local and remote objects (and there is a precedent for the use of such tasks in related HCI research e.g., [19, 20, 24]). Initially two participants were recruited for a pilot study, in order to validate and refine the task. After some minor tweaks to object placement and user instructions, 12 participants (6 pairs) were run through the task. Participants were mostly PhD students from other departments within the University, they were recruited on a self-selected basis responding to an advert and participants were offered £10 compensation. Each pair was gender balanced and all participants had normal or corrected-to-normal vision.

Task overview

The objective was to find and record, using the CamBlend snapshot feature, a set of six different objects provided to each participant via an image sheet. Each participant received a different set of six images making a total of twelve objects to find using the system. In order to enforce collaboration the objects were mixed and placed on either side of the screen. As snapshots can only be recorded of the remote space, individually participants could only record some of the images provided to them. As a consequence participants were required to exchange responsibility for recording snapshots. Finally one object from each image sheet did not exist at all, and participants were encouraged to draw that object on a whiteboard provided.

Participant Instructions

Participants were told that they would need to work together to record all twelve images and that this would involve exchanging images to record. It was made clear that there was no competition element and that participants should take their time. They were encouraged to physically explore their side of the screen, because objects would not be immediately visible and that all the objects could be found within a curtained area. Participants were told that one image from each sheet would need drawing on the whiteboard provided and that the only rule was not to take snapshots of the image sheet. Participants were guided through the task as follows; firstly they were given an instruction sheet detailing all system features and asked to spend some time reading it; they were then led into the study area from separate directions and placed at the CamBlend terminal; participants were given some time to familiarise themselves with the system, the moderator worked with them to make sure all system features were understood and demonstrated in practice; participants were then given the image sheets and the task was explained. After the task was completed pairs of participants were interviewed. In some specific cases where participants

exhibited particular behaviour or problems, this was also an opportunity to query the participant about that.

Environment setup

The system was setup to simulate a two-person teleconference, with two terminals in a partitioned room viewing the opposing space (as described, the studied system does not contain the video encoding necessary for a remote set-up), see Figure 3.

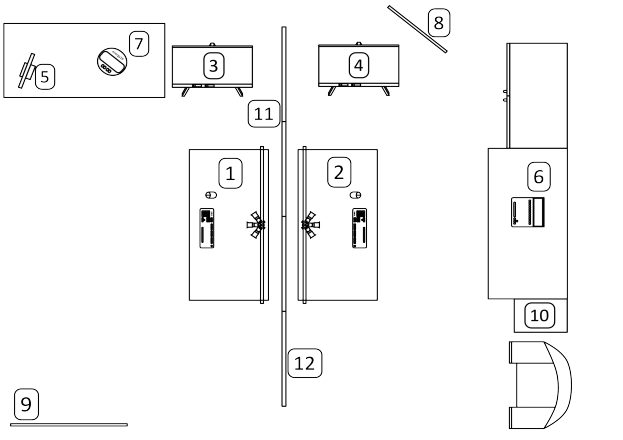


Figure 3. The study area for the task, each space is separated by a non-transparent screen shown.

Each terminal was configured with two focus windows, with the intention that they could permanently focus on something static (the remote participants face) while moving the other around. No suggestions were given to participants on how to use them though.

As well as the objects being searched for, each space was relatively cluttered with general office equipment. This included various computing peripherals, some DIY equipment, some office storage including desk draws and a cupboard. The objects participants were asked to locate were designed to emulate physical properties of objects that might be interesting within an office environment, that is they were either fixed in position, occluded by other objects or the participant themselves, out of camera view (and thus required users to move and 'frame' the object), restricted in their movement by a cable, clearly outside of a normal camera FOV, or needed to be created by drawing on a whiteboard provided. Some of the objects were chosen because they do not have a name, requiring participants to in some way negotiate a shared language to describe them, and others were common objects that can easily be recognised. There were also two monitors behind each user on each side of the screen that cycled through a different set of images. Each monitor was directed towards the CamBlend system so snapshots could be recorded of the contents of the screen.

Methods

To record the study session both the CamBlend screens were recorded via screen capture software, and each participant was recorded via an external fixed position camcorder. The video capture configuration resulted in 4 video feeds to analyse for each group. The 4 video streams were aggregated into a single synchronised high-resolution video stream showing all the recordings together. This video consisted of a 2x2 grid of the source video streams, using the audio track of the video camera with the best audio quality. The analysis of this data uses qualitative video analysis [11], which draws on conversation and interaction analysis techniques [16]. Our primary analytic concern was our participants' practical accomplishment of reciprocal object-focused referencing and how issues of view reconciliation interplay with that accomplishment.

RESULTS

All groups of participants completed the task within 20 minutes, with the fastest group taking just 9 minutes. The moderator who remained in close proximity throughout observed no significant problems with task completion.

Interview data

The questions in a semi-structured interview covered task comprehension, system interface, system feature use and potential problems or improvements. The interviews were recorded via audio and the main points are discussed below. All participants stated that the task was straight forward, that the controls were intuitive and no groups could suggest any improvements to the controls or interface. Most groups said that they enjoyed the task, and that it would have been very difficult without the feedback and pointing tools provided. No groups reported problems with view reconciliation, i.e. reconciling the detail with the overview, and one group reported the opposing problem that they sometimes found it difficult to find the focus windows within the background. Some participants stated that the low resolution provided basic awareness without invading the space, maintaining a level of privacy. At the same time, participants reported feeling uncomfortable using the focus windows to zoom in on faces unless it was clearly necessary. However, they did want to see faces, but wanted it automated both for convenience, and so it was known that they weren't intentionally concentrating on the face. The two participants who didn't like the design of CamBlend both separately described how it felt unfamiliar to have a video stream with differing levels of detail.

Most participants said that screen to physical space transitions were fine, despite evidence during the task of short confusion. Groups described two tactics they used to avoid the need to perform a conversion. One was to rely on the physical space by exploring locally and then physically pointing. The other was to rely exclusively on the system, to '*give over to [the system]*' and ignore your physical surroundings. Also some participants would have liked the

feedback window to be mirrored. Several application areas were brought up, including cross-campus lab sessions, a collaborative storyboarding or documenting tool and as an exploration game. Additionally, improvements were suggested including enhanced snapshot features, privacy controls, better zoom detail and object tracking technology.

Video data

Beyond the interview data, which has a provided an initial overview of system use, the video material provided a more detailed window into intricacies of interaction through and around CamBlend. From this material, it was evident that participants structured the way they handled the task into a number of activities. These were exchanging information on objects left to find (*informing*), exploring to find an object either locally or remotely (*exploring*), communicating the location of objects (*directing*) and framing to record an object (*recording*). The task parameters dictate that each of these activities was of a collaborative nature. To analyse the video it was split temporally into episodes. Those which demonstrated the activities introduced above were transcribed and analysed in detail, resulting in 60 episodes across the six pairs of participants. All names in the following transcript sections have been anonymised. The transcription orthography is based on [11], except that delays are rounded to the nearest half-second, and some notation is omitted, including raising/falling intonation, emphasis and audible inhalations. The original orthography is for short highly detailed transcriptions of around 30s, these changes accommodate for the longer data here.

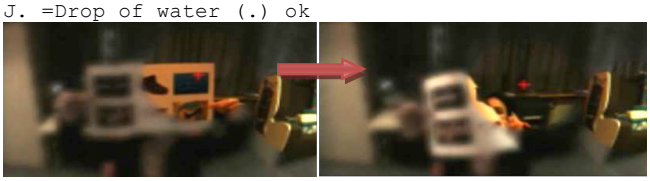
Informing

Although not suggested by the moderator, most participants opted to hold their image sheet up to the camera in order to communicate which images are remaining to find, rather than trying to verbally describe the images in turn. In holding the image sheet up participants would demonstrate one of two tactics. Either they would hold the image sheet still, allowing the remote participant to explore the images. Or sometimes if a particular object was already in discussion, they would attempt to place this image directly over the remote participants focus window, circumventing the remote participants need to operate the system at all. In both cases participants who held an image sheet up demonstrated use of the feedback window to understand remote participants attention area. Often this information was used to query specific images, without the need to describe them. Two episodes are detailed below showing examples of both these techniques.

Session 1, Episide 1

This episode was at the very start of the session, participants are showing each other their entire image sheet with no specific image in mind.

```
L. Right I'll show you
J. °That's the CBI° so you've got a bose speakers
(0.5) er °lets have a° (3.5) can you just pull it
back a bit towards you slightly (1) So we've got a
boot (.) light (.) a globe (0.5) what's this one
here? (1.5)
L. Er:: it's a drop of water=
J. =Drop of water (.) ok
```



Lucy raises her image sheet up to the camera so that John can explore them using his focus window. Initially Lucy holds the sheet too close and John asks it to be moved back. He then moves his focus window over the images one by one, positioning his focus window onto the centre of each image in turn, he pauses over the drop of water and asks '*what's this one here?*'. Lucy then curls the image sheet back where Johns focus window rests and describes the image. Lucy was able to understand immediately which image John meant without the need for any verbal description of the object. In order to understand which image John was querying, Lucy must have been watching the feedback window on her screen from underneath the image sheet as she holds it. This behaviour of watching the feedback window was consistent through all groups.

Session 3, Episode 6

In the following episode Steve requires Rachel to record an image behind him. Steve first needs Rachel to understand which object he has found for her to take a snapshot.

```
S. Oh the bose, sorry on my umm: (0.5) I'll show
you where that is hang on ee:: its behind me: (.)
there we go
S. Yeah so this: (1) this this bose logo (.) there
R. Ah yeah
S. Yep (.) t- that's that's there
R. Ah ok (.) I'll zoom [ (in on that)
S.                     [ on the speaker
R. Sorted (.) so
```

Steve starts by moving Rachel's focus window over to the Bose speaker behind him. Once Rachel's focus window is lined up he shows her which object he requires by holding the image sheet up, he lines it up precisely so the speaker logo is already over Rachel's focus window. Then when he lowers the image sheet Rachel is already looking at the real object. Steve was easily able to convey to Rachel which image he needed, and subsequently where that object was located in the physical environment. Rachel had very little input to the episode, simply confirming her understanding of the situation and then later recording the snapshot.

Exploring

A centre point of the task was actually finding the objects provided to participants via their image sheet. There were three different tactics demonstrated to achieve this. Either participants would look around themselves and explore the space without using the system. They would use the system to explore the opposing space, ignoring the space local to them. Or they would explore their local space using the feedback window, instead of turning around. Participants when appropriate were also combining the tactics. Below are two examples of exploring using the system.

Session 4, Episode 4

Here Emily and James are both exploring the remote space using the system.

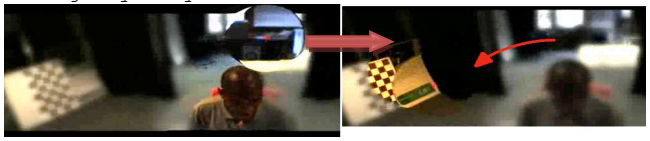
E. Alright (.) around you is there a shoe?
J. A shoe
E. A black sho- a black boot (.) shoe (10)
J. What behind you (or)
E. Th- there isn't (1) I can't see any shoes near me (.) I can't see any shoes near you either
J. No
E. So I'm wondering whether my sh- the shoe is the (.) potential red herring

Here participants are moving their focus windows systematically over the remote environment, zooming into some objects to get more detail as they go. Although some objects were deliberately occluded from the system, this tactic of relying on the system correctly brings them to conclude that the shoe is the extra object that they must draw. This is an example of system reliance described later.

Session 2, Episode 4

Here Harry has found some writing off a poster by directing Jenny's focus window towards it.

H. I've got a board that says supported by CBI voice of business
J. °Ah: ok:°
H. Ah (.) I think its right behind me (1) there (1.5) I think that ones there
J. Ah:: yeah y- y- yeah
H. Is [that it?
J. [Um
J. I- I've got (.) I can see CBI in: the if I magnify it
H. Right yeah yeah



Harry spots the "CBI" writing from the feedback window, without looking behind himself. He then directs Jenny towards the poster by dragging her focus window over to it. When Jenny recognises the object Harry then opts to ask Jenny whether it is the correct image (rather than turning around). It seems here that Harry would rather rely on the system than actually look behind himself at the poster, even to the extent that he asks Jenny to identify the object using the zoom functionality at her disposal.

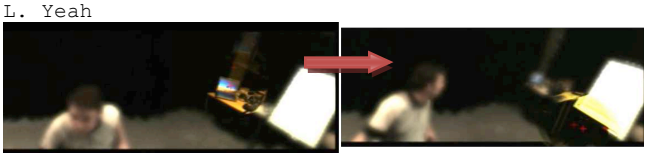
Directing

The task required two different scenarios for directing attention. Either an object was found locally, so participants would need to direct remote attention towards that object. Or vice-versa if an object was found in the remote space (using the system) participants would need to direct a remote users attention to that object. Although it might be possible to avoid one or the other via carefully exchanging responsibilities, all groups demonstrated using both. Depending on how the object was found, the study revealed a multitude of tactics appropriate to either local or remote directing. For instance if a local object is found that is portable, often attention was drawn by simply picking that object up. Additionally, participants sometimes opted to explore their local space using the system rather than physically, so attention could easily be drawn by moving a remote focus window. Below are some examples.

Session 1, Episode 2

Here Lucy believes she has found a shoe on the desk near John, but it is partially occluded and the system doesn't provide enough fidelity for her to recognise. She wants John to look at the desk to check. This is an example of directing remote attention towards a local object.

L. Is that a boot? (3)
J. Where are you looking?
L. On the desk
J. °er::° (1) so that's right left so it's near the board (3) I can't see one on there, there's a bag (.) on the desk (2)
L. Yeah

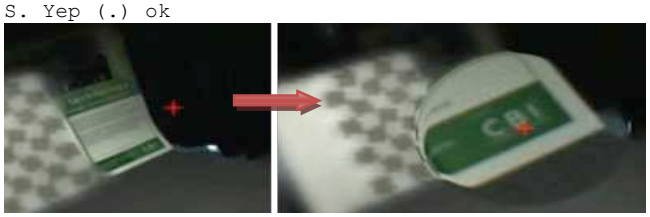


Although John does ask Lucy to describe the location of the shoe "on the desk", he still opts to find it using the feedback window. Here there are clear screen to physical space conversion problems verbalised by John as he calculates the position of the desk. As John states 'so that's right left so' he is turning is head, acting the calculation process. Later he finds the desk and tells Lucy that there is no shoe. This episode also shows that Lucy has implicitly pointed to the desk in the process of attempting to identify the shoe. It's clear that John is using the feedback window in order to identify the shoe location, rather than the verbal description.

Session 5, Episode 2

Sarah has found one of the objects embedded in a poster behind her, so Graham must take a snapshot of it. This is an example of directing remote attention towards local objects.

S. So, I need something from you as well [Graham], so behind me there is this poster, has two green bits, er, I need the bottom one that says CBI (.) the voice of business
S. I think yeah I think you're you're about right no- there yeah (0.5) that should be correct
G. Ok (.) got that
S. Yep (.) ok

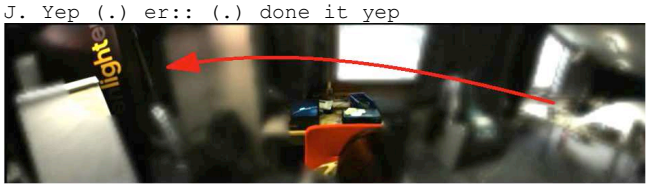


After the initial description Graham understands the general area that Sarah means, so his focus window moves over to the poster and scans around to find the specific writing. Sarah now plays a sort of "hot and cold" game to direct Graham onto the writing, encouraging him when his focus window gets nearer. Graham didn't need to relay information back to Sarah on whether he had located the poster, Sarah was watching the feedback window and thus was aware that Graham was searching the poster area. This is an example of implicitly pointing. Had there been enough resolution for Graham to recognise the poster, he needn't have moved his focus window over to it. Further, recognition via resolution is not sufficient to gain a mutual understanding of which object is correct. Although Graham has received some verbal directions he could only guess exactly which object is in question without confirmation.

Session 4, Episode 2

Emily has spotted one of the objects behind her and directs James over to it in order to take a snapshot. She does so by manipulating James' focus window.

E. E:rm (0.5) so (.) behind me (.) I know where one is ones behind me (.) where I've l- moved my marker where it says light (1.5)
J. Oh yeah
E. Can you take a picture of that?
J. Yep (.) er:: (.) done it yep



Here Emily opts to find the object using the system, even though it might have been easier to turn around and look. After she locates the object, she direct James' attention by manipulating his focus window. James then records a snapshot and verbally confirms that he has done so. As the object was found using the system, it was easy for Emily to manipulate the remote focus window to direct attention, avoiding the need to physically locate the object.

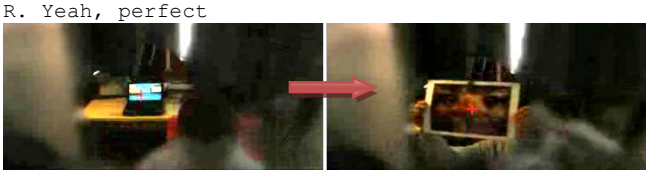
Recording

Recording an object means lining up the focus window over the object and taking a snapshot. If the object was visible and accessible, there were no recorded problems in recording the object. As described, sometimes the object required bringing into the camera view to record, and here participants needed to 'frame' the object to be recorded.

Session 3, Episode 4

Rachel has her focus window already over the monitor behind Steve, she is waiting for a parrot to appear. Meanwhile Steve spots the owl image sheet and brings it into the camera view for Rachel to record.

R. [OK I'll] get that I'll get that when it comes up next
S. [Umm]
S. OK (.) Yeah I've got I've got (1) where's my crosshair (0.5) oh there you go
S. [(like that)] is that in the right place?
R. [°Ah ok°]
R. Yeah, perfect



Steve knew that Rachel was waiting for the parrot on the monitor behind him, and so places the image sheet of the owl directly over her focus window. Rachel doesn't need to move her focus window at all to record both images, she takes a snapshot when Steve holds the image and again later to record the parrot. In waiting for the parrot, Rachel doesn't really intend to point at the monitor, but is implicitly doing so. Steve uses this information to position the owl sheet over her view.

DISCUSSION

The implementation of this focus-in-context video system means that it remains scalable in principle, with multi-point access to the connected spaces available, in contrast to PTZ implementations [24]. Sets of focus windows are then controllable in a distributed fashion, with both the consumer and the server of the views able to move their position. Control remains fast, with no observable latency in moving focus windows over the context background, while this has only been tested with two nodes and four focus areas at present. Although relatively expensive mid-range cameras were used for CamBlend at this stage, much lower cost cameras would principally work with the software infrastructure, resulting in a cost-effective but high-quality video collaboration tool, which can be flexibly deployed. The potential bandwidth savings of the focus-in-context video approach are significant, this test setup uses in the order of 15x fewer pixels in comparison to streaming the entire image. Even though available bandwidth is constantly increasing, such bandwidth savings remain attractive. This is especially so when one considers that many more than three cameras might be assembled resulting in ultra-high resolution live imagery, only sub-sections of which ever need to be streamed, while those sub-sections remain in context of the scene.

As interview and video data have highlighted, CamBlend supported the study task well. Across six pairs the overall task was broken down into the four mentioned activities, for which different strategies were found. Analysis of system use showed that participant conversation tended to transcend the tools with participants talking about the objective at hand. The data collected also allows a brief reflection on some key aspects of interaction.

Reconciling focus and context views

CamBlend provides multiple views into the connected remote space, one context overview and multiple lenses visually imbedded in this overview. The analysis of system use of this focus-in-context video system has shown that this strategy does seem to address the previously reported interactional issues with view reconciliation [9]. It also confirms previous work emphasising the importance of providing legible links between overview and detail views [24]. With CamBlend, those links are established by leaving the detailed view within the context itself. The success of this strategy is evident in the fact that no examples of problems with focus and context view reconciliation can be found in interview or video analysis. To the contrary, some participants identified the integration of detail view into context view as a potential problem, as it was difficult to tell them apart, unless the detail view was zoomed. The visual manipulations on the background image (saturation, de-noising) were partly introduced to avoid this confusion.

Object Referencing

In referencing local or remote objects during the study task, there were two principal observations.

Line of sight

It has been noted that during video collaboration, when making reference to remote resources users often attempt to point 'at screen' into the remote space [15]. The spatial inconsistencies of video break line of sight and thus cause interactional problems. CamBlend does not attempt to repair the video to restore line of sight, in fact displaying the whole panorama accentuates these problems via representation of a much more extreme aspect ratio and FOV. A gesture towards the side of the screen refers to an object sharply to the side of the remote participant, but looks like a gesture pointing roughly forwards.

However, it seems that the flexibility and speed of the focus windows allowed physical 'at-screen' pointing to be all but replaced. This seems to be well represented during the study, where participants never attempted to point 'at-screen' to reference remote objects. By looking into the remote space, participants implicitly indicated their focus of attention, as this was visible in the feedback window. Equally where participants pointed to resources locally, the feedback window enabled this to happen in reverse. This allowed participants to point at artefacts in their physical space. The episode shown in Session 3, Episode 6 demonstrates the success of this mechanism, highlighting the system internal line of sight established in CamBlend. The focus window establishes line of sight between participant and remote artefact. This line of sight is legible and can be acted upon by the remote participant via the feedback window.

Conversion Problems

In addition to a view into the connected remote space, CamBlend also provides a perspective on the local space via the feedback window. All groups demonstrated the process of finding an object physically that was referred to on-screen inside this feedback window. In doing so participants sometimes had problems converting between the physical and screen spaces, shown most clearly inside Session 1, Episode 2. These conversion problems are reflected in similar issues surrounding the motivations for shared visual spaces [6], in that separate interaction spaces trigger delays or confusion in reconciling these views.

During the interviews some participants seemed aware of these problems, and gave one of two tactics for coping. Participants would rely exclusively on either the physical or screen spaces, and try to ignore the other. What is interesting is that the task dictated these tactics would be unlikely to work, via a combination of occluded objects forcing physical exploration, and fixed objects forcing system exploration (even if first found physically a participant would still need to take a snapshot). Further, in some interviews participants were pushed to recall examples of a conversion, successful or not, and they struggled to do so. This suggests that although participants were aware of the difficulty (by recalling a mechanism to cope), they weren't aware that it was a problem. What we suggest is that groups naturally converged into shared tactics for referencing objects, tactics more complex than exclusively screen or physical spaces. This seems to be broadly backed up by the video analysis when all episodes for a group are considered together.

Reflection on application areas

Although the analysis categories (informing, exploring, directing, recording) were chosen to reflect the arrangement of participants activities to complete the study task, here it is briefly shown how each of these categories relate to application areas suggested in the focus groups and beyond. *Informing*, or describing an item to a remote participant is common when establishing a mutual understanding of objects local to one participant, useful for example in the connected classroom scenario supporting creative sessions. *Exploring* is especially useful in observation scenarios where remote users may be unable to help, as for example in the example of connecting to an archaeological dig site. *Directing* remote attention has application areas where attention could be focused on either side of the collaboration, for example the teacher directing attention to a particular student group. Finally *recording* finds

application when the review of data is critical, as in the example of CamBlend as a tool to record dance classes and to review progress. In addition to those, the current system design has also highlighted other potential application areas, such as CCTV, flexible video broadcasting and expert-helper scenarios. We acknowledge that some of these applications would require a substantially different system design and that our study results would not directly be transferrable to those.

CONCLUSION

CamBlend is a lightweight video collaboration tool, which addresses the aforementioned technical problems of scalability, speed and view reconciliation involved in increasing the FOV of video in a novel way. The study of CamBlend highlighted a number of successful examples of object referencing using the pointing tools provided. These included pointing implicitly, explicitly and referencing objects in either local or remote spaces. The problem of screen to physical space transition is identified and a number of tactics, which participants used to cope with the scenario, are highlighted.

ACKNOWLEDGMENTS

We thank all study participants and gratefully acknowledge support from the EPSRC through grants EP/P504252/1 and EP/J006688/1, the Creativity Greenhouse project.

REFERENCES

1. NVidia CUDA. http://www.nvidia.com/cuda.
2. Baudisch, P., Good, N., Bellotti, V., Schraedley, P., Keeping things in context: a comparative evaluation of focus plus context screens, overviews, and zooming, *Proc. CHI '02*, ACM (2002). 259-266
3. Birnholtz, J., Ranjan, A., Balakrishnan, R., Providing Dynamic Visual Information for Collaborative Tasks: Experiments With Automatic Camera Control, *Human-Computer Interaction 2010*, 261-287
4. Cutler, R., Rui, Y., Gupta, A., Cadiz, J.J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z. and Silverberg, S. Distributed meetings: A meeting capture and broadcasting system, in *Proc. ACM Multimedia* (2002)
5. Fussell, S. R., Setlock, L.D., Yang, J., Ou, J., Mauer, E. M., & Kramer, A., Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction 2004,* 273-309
6. Fussell, S.R., Kraut , R.E., and Siegel, J., Coordination of communication: effects of shared visual context on collaborative work, in *Proc. CSCW'00* ACM (2000)
7. Foote, J. and Kimber, D., FlyCam: Practical panoramic video and automatic camera control. in *Multimedia '00* (2000), 487-488.
8. Gaver, W.W., The affordances of media spaces for collaboration, in *Proc. CSCW'92*. ACM (1992), 17-24.
9. Gaver, W.W., Sellen, A., Heath, C and Luff, P., One is not enough: multiple views in a media space, in *Proc. INTERACT '93 and CHI '93.* ACM (1993), 335-341.
10. Heath, C. and Luff, P., Disembodied conduct: communication through video in a multi-media office environment, in *Proc. CHI '91,* ACM (1991), 99-103.
11. Heath, C., Hindmarsh, J., and Luff, P., Video in Qualitative Research, *SAGE.*
12. Heath, C., Knoblauch, H., and Luff, P., Technology and social interaction: the emergence of 'workplace studies'. In *The British Journal of Sociology,* (2000), 299-320.
13. Hindmarsh, J., Fraser, M., Heath, C., Benford, S and Greenhalgh, C, Object-focused interaction in collaborative virtual environments. *TOCHI '00*, ACM (2000), 477-509.
14. Hinds, P.J and Keisler, S. Distributed Work: *MIT Press.*
15. Isaacs, E.A., and Tang, J.C., What video can and cannot do for collaboration: A case study. *Proc. MULTIMEDIA '93* (1993), ACM, 199-206.
16. Jordan, B and Henderson, A., Interaction Analysis: Foundations and Practice. in *Journal of Learning Sciences,* PARC (1994), 39-103
17. Keahey, T.A., and Robertson, E.L., Techniques for non-linear magnification transformations. In *Proc. Inf. Vis. '96*, IEEE (1996), 38-45
18. Kuzuoka, H., Oyama, S., Yamazaki, K., Suzuki, K and Mitsuishi, M., GestureMan: a mobile robot that embodies a remote instructor's actions, in *Proc. CSCW '00.* ACM (2000), 155-162
19. Luff, P., Yamashita, N., Kuzuoka, H and Heath, C, Hands on hitchcock: embodied reference to a moving scene, In *Proc. CHI '11,* ACM (2011), 43-52
20. Luff, P., Heath, C., Kuzuoka, H., Hindmarsh, J., Yamazaki, K and Oyama, S, Fractured Ecologies: Creating Environments for Collaboration. In *Human-Computer Interaction 2003,* 51-84
21. Mantei, M.M., Baecher, R.M., Sellen, A.J., Buxton, W.A.S., Milligan, T and Wellman, B., Experiences in the Use of a Media Space. In *CHI '91.* ACM (1991)
22. O'hara, K., Kjeldskov, J., and Paay, J., Blended interaction spaces for distributed team collaboration. In *TOCHI '11*. ACM (2011)
23. Ranjan, A., Birnholtz, J, P., and Balakrishnan, R., An exploratory analysis of partner action and camera control In a video-mediated collaborative task. *Proc. CSCW '06,* ACM (2006), 403-412.
24. Yamaashi, K., Cooperstock, J, R., Narine, T., Buxton, W., Beating the limitations of camera-monitor mediated telepresence with extra eyes, in *Proc. CHI '06.* ACM (1996), 50-5