CLUE WG A. Romanow
Internet-Draft Cisco
Intended status: Informational S. Botzko
Expires: November 26, 2011 M. Duckworth
Polycom
R. Even
Huawei Technologies
T. Eubanks
Iformata Communications
May 25, 2011

Use Cases for Telepresence Multi-streams
draft-romanow-clue-telepresence-use-cases-02.txt

Abstract

Telepresence conferencing systems seek to create the sense of really being present. A number of techniques for handling audio and video streams are used to create this experience. When these techniques are not similar, interoperability between different systems is difficult at best, and often not possible. Conveying information about the relationships between multiple streams of media would allow senders and receivers to make choices to allow telepresence systems to interwork. This memo describes the most typical and important use cases for sending multiple streams in a telepresence conference.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

Telepresence applications try to provide a "being there" experience for conversational video conferencing. Often this telepresence application is described as "immersive telepresence" in order to distinguish it from traditional video conferencing, and from other forms of remote presence not related to conversational video conferencing, such as avatars and robots. The salient characteristics of telepresence are often described as: full-sized, immersive video, preserving interpersonal interaction and allowing non-verbal communication.

Although telepresence systems are based on open standards such as RTP [RFC3550], SIP [RFC3261] , H.264, and the H.323 suite of protocols, they cannot easily interoperate with each other without operator assistance and expensive additional equipment which translates from one vendor to another. A standard way of describing the multiple streams constituting the media flows and the fundamental aspects of their behavior, would allow telepresence systems to interwork.

This draft presents a set of use cases describing typical scenarios. Requirements will be derived from these use cases in a separate document. The use cases are described from the viewpoint of the users. They are illustrative of the user experience that needs to be supported. It is possible to implement these use cases in a variety of different ways.

Many different scenarios need to be supported. Our strategy in this document is to describe in detail the most common and basic use cases. These will cover most of the requirements. Additional scenarios that bring new features and requirements will be added.

We look at telepresence conferences that are point-to-point and multipoint. In some settings, the number of displays is similar at all sites, in others, the number of displays differs at different sites. Both cases are considered. Also included is a use case describing display of presentation or content.

The document structure is as follows: Section 2 presents the document terminology, Section 3 gives an overview of the scenarios, and Section 4 describes use cases.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Telepresence Scenarios Overview

This section describes the general characteristics of the use cases and what the scenarios are intended to show. The typical setting is a business conference, which was the initial focus of telepresence. Recently consumer products are also being developed. We specifically do not include in our scenarios the infrastructure aspects of telepresence, such as room construction, layout and decoration.

Telepresence systems are typically composed of one or more video cameras and encoders and one or more display monitors of large size (around 60"). Microphones pick up sound and audio codec(s)produce one or more audio streams. The cameras used to present the telepresence users we will call participant cameras (and likewise for displays). There may also be other cameras, such as for document display. These will be referred to as presentation or content cameras, which generally have different formats, aspect ratios, and frame rates from the participant cameras. The presentation videos may be shown on participant screen, or on auxiliary display screens. A user’s computer may also serve as a virtual content camera, generating an animation or playing back a video for display to the remote participants.

We describe such a telepresence system as sending M video streams, N audio streams, and D content streams to the remote system(s). (Note that the number of audio streams is generally not the same as the number of video streams.)

The fundamental parameters describing today’s typical telepresence scenario include:

1. The number of participating sites

2. The number of visible seats at a site

3. The number of cameras

4. The number of audio channels

5. The screen size

6. The display capabilities - such as resolution, frame rate, aspect ratio

7. The arrangement of the displays in relation to each other

8. Similar or dissimilar number of primary screens at all sites

9. Type and number of presentation displays

10. Multipoint conference display strategies - for example, the camera-to-display mappings may be static or dynamic

11. The camera viewpoint

12. The cameras fields of view and how they do or do not overlap

The basic features that give telepresence its distinctive characteristics are implemented in disparate ways in different systems. Currently Telepresence systems from diverse vendors interoperate to some extent, but this is not supported in a standards based fashion. Interworking requires that translation and transcoding devices be included in the architecture. Such devices increase latency, reducing the quality of interpersonal interaction. Use of these devices is often not automatic; it frequently requires substantial manual configuration and a detailed understanding of the nature of underlying audio and video streams. This state of affairs is not acceptable for the continued growth of telepresence - we believe telepresence systems should have the same ease of interoperability as do telephones.

There is no agreed upon way to adequately describe the semantics of how streams of various media types relate to each other. Without a standard for stream semantics to describe the particular roles and activities of each stream in the conference, interoperability is cumbersome at best.

In a multiple screen conference, the video and audio streams sent from remote participants must be understood by receivers so that they can be presented in a coherent and life-like manner. This includes the ability to present remote participants at their true size for their apparent distance, while maintaining correct eye contact, gesticular cues, and simultaneously providing a spatial audio sound stage that is consistent with the video presentation.

The receiving device that decides how to display incoming information needs to understand a number of variables such as the spatial position of the speaker, the field of view of the cameras; the camera zoom; which media stream is related to each of the displays; etc. It is not simply that individual streams must be adequately described, to a large extent this already exists, but rather that the semantics of the relationships between the streams must be communicated. Note that all of this is still required even if the basic aspects of the streams, such as the bit rate, frame rate, and aspect ratio, are known. Thus, this problem has aspects considerably beyond those encountered in interoperation of single-node video conferencing units.

4. Use Case Scenarios

Our development of use cases is staged, initially focusing on what is currently typical and important. Use cases that add future or more specialized features will be added later as needed. Also, there are a number of possible variants for these use cases, for example, the audio supported may differ at the end points (such as mono or stereo versus surround sound), etc.

The use cases here are intended to be hierarchical, in that the earlier use cases describe basics of telepresence that will also be used by later use cases.

Many of these systems offer a full conference room solution where local participants sit on one side of a table and remote participants are displayed as if they are sitting on the other side of the table. The cameras and screens are typically arranged to provide a panoramic (left to right from the local user view point) view of the remote room.

The sense of immersion and non-verbal communication is fostered by a number of technical features, such as:

1. Good eye contact, which is achieved by careful placement of participants, cameras and screens.

2. Camera field of view and screen sizes are matched so that the images of the remote room appear to be full size.

3. The left side of each room is presented on the right display at the far end; similarly the right side of the room is presented on the left display. The effect of this is that participants of each site appear to be sitting across the table from each other. If two participants on the same site glance at each other, all participants can observe it. Likewise, if a participant on one site gestures to a participant on the other site, all participants observe the gesture itself and the participants it includes.

4.1. Point to point meeting: symmetric

In this case each of the two sites has an identical number of screens, with cameras having fixed fields of view, and one camera for each screen. The sound type is the same at each end. As an example, there could be 3 cameras and 3 screens in each room, with stereo sound being sent and received at each end.

The important thing here is that each of the 2 sites has the same number of screens. Each screen is paired with a corresponding camera. Each camera / screen pair is typically connected to a separate codec, producing a video encoded stream for transmission to the remote site, and receiving a similarly encoded stream from the remote site.

Each system has one or multiple microphones for capturing audio. In some cases, stereophonic microphones are employed. In other systems, a microphone may be placed in front of each participant (or pair of participants). In typical systems all the microphones are connected to a single codec that sends and receives the audio streams as either stereo or surround sound. The number of microphones and the number of audio channels are often not the same as the number of cameras. Also the number of microphones is often not the same as the number of loudspeakers.

The audio may be transmitted as multi-channel (stereo/surround sound) or as distinct and separate monophonic streams. Audio levels should be matched, so the sound levels at both sites are identical. Loudspeaker and microphone placements are chosen so that the sound "stage" (orientation of apparent audio sources) is coordinated with the video. That is, if a participant on one site speaks, the participants at the remote site perceive her voice as originating from her visual image. In order to accomplish this, the audio needs to be mapped at the received site in the same fashion as the video. That is, audio received from the right side of the room needs to be output from loudspeaker(s) on the left side at the remote site, and vice versa.

4.2. Point to point meeting: asymmetric

In this case, each site has a different number of screens and cameras than the other site. The important characteristic of this scenario is that the number of displays is different between the two sites. This creates challenges which are handled differently by different telepresence systems.

This use case builds on the basic scenario of 3 screens to 3 screens. Here, we use the common case of 3 screens and 3 cameras at one site, and 1 screen and 1 camera at the other site, connected by a point to point call. The display sizes and camera fields of view at both sites are basically similar, such that each camera view is designed to show two people sitting side by side. Thus the 1 screen room has up to 2 people seated at the table, while the 3 screen room may have up to 6 people at the table.

The basic considerations of defining left and right and indicating relative placement of the multiple audio and video streams are the same as in the 3-3 use case. However, handling the mismatch between the two sites of the number of displays and cameras requires more complicated maneuvers.

For the video sent from the 1 camera room to the 3 screen room, usually what is done is to simply use 1 of the 3 displays and keep the second and third displays inactive, or put up the date, for example. This would maintain the "full size" image of the remote side.

For the other direction, the 3 camera room sending video to the 1 screen room, there are more complicated variations to consider. Here are several possible ways in which the video streams can be handled.

1. The 1 screen system might simply show only 1 of the 3 camera images, since the receiving side has only 1 screen. Two people are seen at full size, but 4 people are not seen at all. The choice of which 1 of the 3 streams to display could be fixed, or could be selected by the users. It could also be made automatically based on who is speaking in the 3 screen room, such that the people in the 1 screen room always see the person who is speaking. If the automatic selection is done at the sender, the transmission of streams that are not displayed could be suppressed, which would avoid wasting bandwidth.

2. The 1 screen system might be capable of receiving and decoding all 3 streams from all 3 cameras. The 1 screen system could then compose the 3 streams into 1 local image for display on the single screen. All six people would be seen, but smaller than full size. This could be done in conjunction with reducing the image resolution of the streams, such that encode/decode resources and bandwidth are not wasted on streams that will be downsized for display anyway.

3. The 3 screen system might be capable of including all 6 people in a single stream to send to the 1 screen system. For example, it could use PTZ (Pan Tilt Zoom) cameras to physically adjust the cameras such that 1 camera captures the whole room of six people. Or it could recompose the 3 camera images into 1 encoded stream to send to the remote site. These variations also show all six people, but at a reduced size.

4. Or, there could be a combination of these approaches, such as simultaneously showing the speaker in full size with a composite of all the 6 participants in smaller size.

The receiving telepresence system needs to have information about the content of the streams it receives to make any of these decisions. If the systems are capable of supporting more than one strategy, there needs to be some negotiation between the two sites to figure out which of the possible variations they will use in a specific point to point call.

4.3. Multipoint meeting

In a multipoint telepresence conference, there are more than two sites participating. Additional complexity is required to enable media streams from each participant to show up on the displays of the other participants.

Clearly, there are a great number of topologies that can be used to display the streams from multiple sites participating in a conference.

One major objective for telepresence is to be able to preserve the "Being there" user experience. However, in multi-site conferences it is often (in fact usually) not possible to simultaneously provide full size video, eye contact, common perception of gestures and gaze by all participants. Several policies can be used for stream distribution and display: all provide good results but they all make different compromises.

One common policy is called site switching. Let’s say the speaker is at site A and everyone else is at a "remote" site. When the room at site A shown, all the camera images from site A are forwarded to the remote sites. Therefore at each receiving remote site, all the screens display camera images from site A. This can be used to preserve full size image display, and also provide full visual context of the displayed far end, site A. In site switching, there is a fixed relation between the cameras in each room and the displays in remote rooms. The room or participants being shown is switched from time to time based on who is speaking or by manual control, e.g., from site A to site B.

Segment switching is another policy choice. Still using site A as where the speaker is, and "remote" to refer to all the other sites, in segment switching, rather than sending all the images from site A, only the speaker at site A is shown. The camera images of the current speaker and previous speakers (if any) are forwarded to the other sites in the conference. Therefore the screens in each site are usually displaying images from different remote sites - the current speaker at site A and the previous ones. This strategy can be used to preserve full size image display, and also capture the non-verbal communication between the speakers. In segment switching, the display depends on the activity in the remote rooms - generally, but not necessarily based on audio / speech detection).

A third possibility is to reduce the image size so that multiple camera views can be composited onto one or more screens. This does not preserve full size image display, but provides the most visual context (since more sites or segments can be seen). Typically in this case the display mapping is static, i.e., each part of each room is shown in the same location on the display screens throughout the conference.

Other policies and combinations are also possible. For example, there can be a static display of all screens from all remote rooms, with part or all of one screen being used to show the current speaker at full size.

4.4. Presentation

In addition to the video and audio streams showing the participants, additional streams are used for presentations.

In systems available today, generally only one additional video stream is available for presentations. Often this presentation stream is half-duplex in nature, with presenters taking turns. The presentation video may be captured from a PC screen, or it may come from a multimedia source such as a document camera, camcorder or a DVD. In a multipoint meeting, the presentation streams for the currently active presentation are always distributed to all sites in the meeting, so that the presentations are viewed by all.

Some systems display the presentation video on a screen that is mounted either above or below the three participant screens. Other systems provide monitors on the conference table for observing presentations. If multiple presentation monitors are used, they generally display identical content. There is considerable variation in the placement, number, and size or presentation displays.

In some systems presentation audio is pre-mixed with the room audio. In others, a separate presentation audio stream is provided (if the presentation includes audio).

In H.323 systems, H.239 is typically used to control the video presentation stream. In SIP systems, similar control mechanisms can be provided using BFCP [RFC4582] for presentation token. These mechanisms are suitable for managing a single presentation stream.

Although today’s systems remain limited to a single video presentation stream, there are obvious uses for multiple presentation streams.

1. Frequently the meeting convener is following a meeting agenda, and it is useful for her to be able to show that agenda to all participants during the meeting. Other participants at various remote sites are able to make presentations during the meeting, with the presenters taking turns. The presentations and the agenda are both shown, either on separate displays, or perhaps re-scaled and shown on a single display.

2. A single multimedia presentation can itself include multiple video streams that should be shown together. For instance, a presenter may be discussing the fairness of media coverage. In addition to slides which support the presenter’s conclusions, she also has video excerpts from various news programs which she shows to illustrate her findings. She uses a DVD player for the video excerpts so that she can pause and reposition the video as needed. Another example is an educator who is presenting a multi-screen slide show. This show requires that the placement of the images on the multiple displays at each site be consistent.

There are many other examples where multiple presentation streams are useful.

4.5. Heterogeneous Systems

It is common in meeting scenarios for people to join the conference from a variety of environments, using different types of endpoint devices. In a multi-screen immersive telepresence conference may include someone on a PC-based video conferencing system, a participant calling in by phone, and (soon) someone on a handheld device.

What experience/view will each of these devices have?

Some may be able to handle multiple streams and others can handle only a single stream. (We are not here talking about legacy systems, but rather systems built to participate in such a conference, although they are single stream only.) In a single video stream , the stream may contain one or more compositions depending on the available screen space on the device. In most cases a transcoding intermediate device will be relied upon to produce a single stream, perhaps with some kind of continuous presence.

Bit rates will vary - the handheld and phone having lower bit rates than PC and multi-screen systems.

Layout is accomplished according to different policies. For example, a handheld and PC may receive the active speaker stream. The decision can either be made explicitly by the receiver or by the sender if it can receive some kind of rendering hint. The same is true for audio -- i. e., that it receives a mixed stream or a number of the loudest speakers if mixing is not available in the network.

For the software conferencing participant, the user’s experience depends on the application. It could be single stream, similar to a handheld but with a bigger screen. Or, it could be multiple streams, similar to an immersive but with a smaller screen. Control for manipulation of streams can be local in the software application, or in another location and sent to the application over the network.

The handheld device is the most extreme. How will that participant be viewed and heard? it should be an equal participant, though the bandwidth will be significantly less than an immersive system. A receiver may choose to display output coming from a handheld differently based on the resolution, but that would be the case with any low resolution video stream, e. g., from a powerful PC on a bad network.

The handheld will send and receive a single video stream, which could be a composite or a subset of the conference. The handheld could say what it wants or could accept whatever the sender (conference server or sending endpoint) thinks is best. The handheld will have to signal any actions it wants to take the same way that immersive signals.

4.6. Multipoint Education Usage

The importance of this example is that the multiple video streams are not used to create an immersive conferencing experience with panoramic views at all the site. Instead the multiple streams are dynamically used to enable full participation of remote students in a university class. In some instances the same video stream is displayed on multiple displays in the room, in other instances an available stream is not displayed at all.

The main site is a university auditorium which is equipped with three cameras. One camera is focused on the professor at the podium. A second camera is mounted on the wall behind the professor and captures the class in its entirety. The third camera is co-located with the second, and is designed to capture a close up view of a questioner in the audience. It automatically zooms in on that student using sound localization.

Although the auditorium is equipped with three cameras, it is only equipped with two screens. One is a large screen located at the front so that the class can see it. The other is located at the rear so the professor can see it. When someone asks a question, the front screen shows the questioner. Otherwise it shows the professor (ensuring everyone can easily see her).

The remote sites are typical immersive telepresence room with three camera/screen pairs.

All remote sites display the professor on the center screen at full size. A second screen shows the entire classroom view when the professor is speaking. However, when a student asks a question, the second screen shows the close up view of the student at full size. Sometimes the student is in the auditorium; sometimes the speaking student is at another remote site. The remote systems never display the students that are actually in that room.

If someone at the remote site asks a question, then the screen in the auditorium will show the remote student at full size (as if they were present in the auditorium itself). The display in the rear also shows this questioner, allowing the professor to see and respond to the student without needing to turn her back on the main class.

When no one is asking a question, the screen in the rear briefly shows a full-room view of each remote site in turn, allowing the professor to monitor the entire class (remote and local students). The professor can also use a control on the podium to see a particular site - she can choose either a full-room view or a single camera view.

Realization of this use case does not require any negotiation between the participating sites. Endpoint devices (and an MCU if present) - need to know who is speaking and what video stream includes the view of that speaker. The remote systems need some knowledge of which stream should be placed in the center. The ability of the professor to see specific sites (or for the system to show all the sites in turn) would also require the auditorium system to know what sites are available, and to be able to request a particular view of any site. Bandwidth is optimized if video that is not being shown at a particular site is not distributed to that site.

5. Acknowledgements

The draft has benefitted from input from a number of people including Alex Eleftheriadis, Tommy Andre Nyquist, Mark Gorzynski, Charles Eckel, Nermeen Ismail, Mary Barnes, Pascal Buhler, Jim Cole.

6. IANA Considerations

This document contains no IANA considerations.

7. Security Considerations

While there are likely to be security considerations for any solution for telepresence interoperability, this document has no security considerations.

8. Informative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC4582] Camarillo, G., Ott, J., and K. Drage, "The Binary Floor Control Protocol (BFCP)", RFC 4582, November 2006.

Authors' Addresses

Allyn Romanow
Cisco
San Jose, CA 95134
US

Email: allyn@cisco.com

Stephen Botzko
Polycom
Andover, MA 01810
US

Email: stephen.botzko@polycom.com

Mark Duckworth
Polycom
Andover, MA 01810
US

Email: mark.duckworth@polycom.com

Roni Even
Huawei Technologies
Tel Aviv,
Israel

Email: even.roni@huawei.com

Marshall Eubanks
Iformata Communications
Dayton, Ohio 45402
US

Email: marshall.eubanks@ilformata.com