Enabling 'Togetherness' in High-Quality Domestic Video Conferencing

[1] Ian Kegel, [2] Pablo Cesar, [2] Jack Jansen, [2] Dick Bulterman, [1] Tim Stevens,
[3] Joke Kort, [4] Nikolaus Färber

[1]BT Research & Technology
Adastral Park
Martlesham Heath
Ipswich IP5 3RE, UK

[2]CWI: Centrum Wiskunde
& Informatica
Science Park 123
1098 XG Amsterdam
Netherlands

[3]TNO ICT
PO Box 15000
9700 CD Groningen
Netherlands

[4]Fraunhofer IIS
Am Wolfsmantel 33
91058 Erlangen
Germany

ian.c.kegel@bt.com, p.s.cesar@cwi.nl, jack.jansen@cwi.nl, dick.bulterman@cwi.nl, tim.s.stevens@bt.com,
joke.kort@tno.nl, nikolaus.faerber@iis.fraunhofer.de

ABSTRACT

Low-cost video conferencing systems have provided an existence proof for the value of video communication in a home setting. At the same time, current systems have a number of fundamental limitations that inhibit more general social interactions among multiple groups of participants. In our work, we describe the development, implementation and evaluation of a domestic video conferencing system that is geared to providing true 'togetherness' among conference participants. We show that such interactions require sophisticated support for high-quality audiovisual presentation, and processing support for person identification and localisation. In this paper, we describe user requirements for effective interpersonal interaction. We then report on a system that implements these requirements. We conclude with a systems and user evaluation of this work. We present results that show that participants in a video conference can be made feel as 'together' as collocated players of a board game.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems – *Human factors*. H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *Audio, Video*.

Keywords

video conferencing, social communication, interaction, visual composition, togetherness

1. INTRODUCTION

In 1910, the artist Villemard created a series of 24 postcards that illustrated how technology would influence life 90 years later, in 2000. Most postcards show flying bicycles and other intricate forms of public transport. In one picture, however, it is not the public at large being transported from one place to another, but a single person (see Figure 1). Sitting in the comfort of his own home, a gentleman is able to interact remotely with his wife (or mistress). A large, lifelike display helps give a feeling of remote presence. Judging from the microphone and the phonograph-like loudspeaker, it is clear that both audio and visual information play important roles in this inter-personal communication.

MM'12, October 29–November 2, 2012, Nara, Japan.

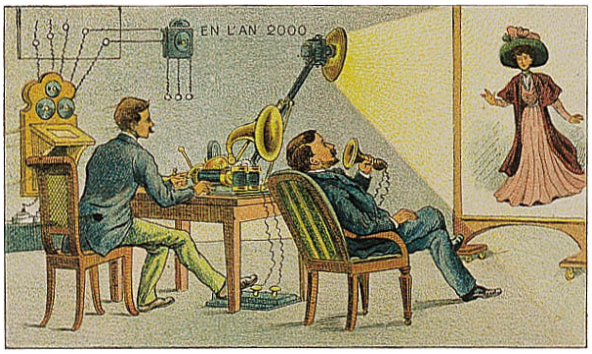


Figure 1: Remote interpersonal communication, as envisioned by Villemard

There is a lot of familiar technology in this picture: we see a network infrastructure, active content filtering (evidenced by the fact that the woman's background context has been removed) and even appropriate furniture. All of this technology is important, but so is the faithful assistant of the couple who is orchestrating the communication between them. Even in 1910, it was clear that microphones, cameras and networks alone were not enough to support intimate interactions.

Figure 2: Remote interpersonal communication, as instantiated by Skype

Part of Villemard's vision has become reality. Although a full 10 years later than Villemard thought, it is now fairly commonplace to have remote audio and visual conferences between people in home settings. The cameras are smaller and the microphones more

subtle, but - as we can see in the image of a Skype-style desktop (see Figure 2) - the naturalness of the communication among the parties still has a way to go before reaching the ideal of Villemard's drawing. With Skype, nobody seems to know where to look, and nobody appears to be 'together'.This paper reports our efforts in supporting high-quality interpersonal communication for domestic videoconferences involving multiple participants at each end. Our goal is to better understand how a multi-camera home environment can help to make communication and engagement easier between groups of people separated in space – a concept we call *togetherness*. We particularly focus on the processes of data capture, encoding and composition, since they play a key role in supporting social communications. These processes make it possible for content originating from multiple sources to be presented at each location in a dynamic and flexible way using embedded control within a system (much like the faithful assistant would do in Villemard's vision).

The specific contributions discussed in this paper are:

- The evaluation of a prototype which supports social video conferencing based around a shared activity, carried out between multiple people at each location by using multiple cameras and dynamic multimedia composition.
- A technical implementation which uniquely combines low-delay, high quality audiovisual communication (including high definition video and multi-channel audio) with dynamic content from a shared application.
- An interaction architecture which provides the ability to dynamically modify audiovisual streams so that the movements of participants, verbal and non-verbal interactions and changes in a shared activity can be presented effectively.
- Application-level support for aesthetically integrating content streams and end-user interaction (such as game play)

This paper describes a system that was used for a series of interactive gaming experiments. It contains an evaluation of technical aspects of our work and the results of a user study on the impact of this technology on a feeling of togetherness among participants. Using objective and subjective measurements, we conclude that high-quality communication not only depends on the technology, but as well on how well social interactions are supported.

The paper begins by summarising related work in Section 2. Section 3 discusses requirements for supporting interaction rituals in remote interpersonal communication. Section 4 describes our technical implementation, focusing on the key data capture, encoding and composition features of the system. Sections 5 and 6 report the results of a number of trials evaluating the system from both a technology and a human perspective. Finally, Section 7 discusses the lessons learned during the process of implementing the system.

2. RELATED WORK

Domestic video conferencing is becoming commonplace, with Skype providing a convincing existence proof on the viability of home video communication. Still, users encounter many limitations with existing technology. Recent studies on human factors have identified common restrictions when using Skype at home. Some of them relate to performance: "Families frequently encounter technical difficulties even after the call is established: unreliable Internet connections, microphones with feedback, video lag or visual artifacts, frozen screens, and crashed applications were all common" [1]. Other restrictions refer to functionality: "The systems used in the homes we observed were often used by multiple people… This suggests a need to develop a home appliance for multiparty viewing and use" [14]. These results corroborate our own research into user requirements [23].

A key assumption within our research is the need to bound social video conferencing with a shared activity. Kirk et al. concluded that "… there were also times when it was clearly important that video could be meshed with other activities as necessary" [14]. Social games such as Mafia [2] provide another example, in which users value the ability to perform a shared activity together with remote parties. It has been shown that *interaction rituals* are important for maintaining social relationships and for building social cohesion and social identity [8]. It has been noted [5][11] that an important part of these rituals is performing mutual activities. It has also been suggested [5] that interaction rituals which make use of the full expressive capacity of human beings will make a stronger impact than those only based on language.

We are seeking to achieve a form of communication that is much less limited, in terms of its embodied interaction, than forms of mediated communication that are popularised today [7], especially by providing a focus on groups and not individuals as actors [17].

Our prototype introduces the dynamic composition of audiovisual streams and content. This functionality has been identified (although not implemented) in other works, highlighting the importance of manipulating and managing components within a set of video streams [10]. Studies on video-mediated free play between children found that different kinds of views led to different types of play [22], while other experiments demonstrate that good framing techniques improve social communication [18], and provide more vivid recorded lectures [16]. More recently, evaluations of remote game playing have shown that framing techniques can improve the effectiveness of the participants [12].

Visual Composition is an architectural block which is not usually present in video chat applications. In video chat, content stream manipulations happen after capture (e.g. effects in Apple's iChat application), while more immersive systems require complex content stream manipulations and camera control [4][19] [21]. In our research, we have applied the latter techniques for improving social video conferencing. Unlike previous research and commercial systems (e.g. from Cisco, Polycom or Lifesize), our composition component is reactive to the social interactions of the participants (e.g. non-verbal cues) and the shared activity (e.g. turn taking in a shared game).

In previous work [13], we reported on experiments to determine the approximate end-to-end delay of both commercial video conferencing systems and video chat applications. These experiments suggest delay figures around 300ms for commercial systems and also for Skype, and less than 200ms for Apple's iChat application – although both video chat applications were using 4CIF video resolution rather than 720p HD. In Section 5, this paper reports on more rigorous experiments which verify that delay figures for typical commercial systems are significantly higher than measured on our prototype system at comparable video resolutions.

3. SUPPORTING INTERACTION RITUALS

Traditionally, user requirements for domestic video conferencing have been defined based on performance measures such as video quality and latency. The requirements have typically focused on

the low-level transfer of communication bits. In contrast, our work has focused on a broader understanding of high-level interpersonal communication.

In order to better understand these requirements, we conducted interviews within 16 families across four countries (U.K., Sweden, Netherlands, and Germany) [13][23]. The families consisted of adults, as well as children aged between about 6 and 25. Based on the interviews, we concluded that the use of video communication (particularly Skype) was not found compelling because of the technology. Users complained about the quality of the video and about the de-synchronisation between audio and video. Results of the interviews allowed us to identify a number of performance requirements: multiple cameras to support the framing of people, high definition video for fitting groups of people in front of the camera(s), high quality multi-channel audio that allows for speaker identification, and low delay audiovisual transmission.

More interestingly, in the interviews playful activities and games were considered as the common way families interacted in person. Some of the interviewees agreed that they did not see the games as an end in themselves, but tended to play them rather as a convenient excuse for getting together physically. This finding imposed a functional requirement for enabling high-quality communications: video communication should be bounded and reactive to social activities.

Based on these results, we designed and implemented a system that supported playful activities between multiple people at remote locations. In particular, we focused on a shared game, *Space Alert*, as a typical activity for a family gathering. Since our goal was in gaining insights into participant togetherness, we wanted to mimic a relatively common living room environment at each location where the game was to be played.

In the first iteration of our system we designed the shared game to closely replicate the physical board game by which it was inspired. The participants used a conventional TV screen predominantly for video communication, while at the same time a touchscreen provided an interactive representation of the game board which was synchronised between locations. This first prototype worked well and enabled us to obtain and report early results on the architecture for dynamic composition of audiovisual streams [13]. But it also showed that the two separate screen presentations (TV and touchscreen) did not support communication very well. Participants tended to focus most of their time on the game board to the detriment of the composed video presentation.

In the second iteration of our system, the focus of this paper, the Space Alert game was adapted away from the traditional board game metaphor with the aim of providing a single focus for participants' attention, less complex game rules, and most importantly to enable subjective evaluation of a new concept: the tight integration of both audiovisual communication, game content and interaction. In this new design, the participants use only the TV screen both for social communication and playing the game. Playing cards embedded with RFID tags are used to control chance aspects of the game, while the participants use their own bodies (via a Microsoft Kinect 3D motion sensor) as the interface to completing a series of mini-games. The game is intrinsically cooperative: players in different locations must collaborate to achieve a common goal in each mini-game – for example collectively steering a ship through an asteroid field – by taking different individual roles which require communication. Figure 3 shows example screenshots and photos of this shared game in practice. Playing a game together like this is a valuable way of fostering strong ties. This is typically an immersive and emotional experience (a ritual) that creates shared memories and thus enhances togetherness in the long term.

(a) Choosing a planet to explore (view of room A)

(b) Choosing a planet to explore (view of room B)

(c) (d) Playing a mini-game (rooms A and B)

Figure 3: Playing the Space Alert shared game. Images (a) and (b) show different presentations of the game at the same time in two locations. Images (c) and (d) show players interacting co-operatively during the game.

A critical component of this shared game is the presentation of the game's features and the effective display of interactions among participants. Composition in togetherness-based applications is inherently dynamic: the content will need to change as the focus of the activity (and the roles of the participants) develops. For example, positioning cameras to allow people to address the TV screen, and hence their friends, from different parts of the room is more useful than a camera that covers the whole space of the room. For collaborative, cooperative games, the game and visual elements should be composited on the same screen to encourage eye contact and to enhance the value derived from the communication. Seeing your partner in a natural, non-monotonic manner seems essential for this activity.

4. SYSTEM DESIGN AND IMPLEMENTATION

In Section 2 we explained how existing video conferencing and video chat systems are limited in their ability to effectively

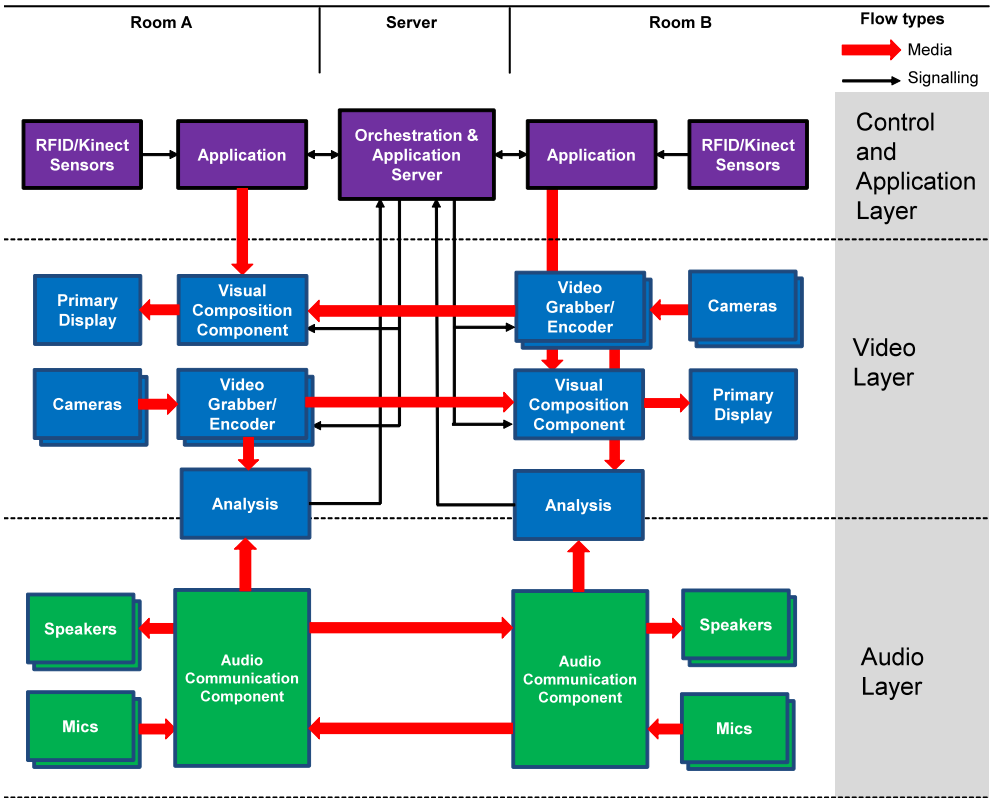


Figure 4: Overall system architecture

address groups and to provide flexible, dynamic composition of both live streams and content related to a shared application. In Section 3 we described how these limitations were reflected by interviews with typical families, and how through two iterations we have created a shared game which is tightly integrated with audiovisual communication. We therefore needed to develop a technical system support the Space Alert experience and to help us gain insights into togetherness in social interactions between groups. This section describes that system, highlighting why its features are important to the achievement of togetherness.

As discussed in Section 3, the system is designed to support a game shared between different locations. This involves rich communication (video and speech), together with a shared application (the game), which can be enjoyed by multiple people, at multiple ends, captured by multiple cameras. The design must be capable of handling both real time and recorded streams of both audio and video at each location independently. In addition the system must be able to intelligently decide how to compose the video and audio outputs at each location. This intelligent decision-making process, which we call *orchestration*, is partly based on analysis of audio and video signals captured in each location.

Figure 4 shows the overall architecture that spans three abstraction layers. The Control and Application layer contains components that could be proprietary (e.g. the games) and those responsible for instructing video composition (*orchestration*). The Video Layer contains all the video processing components. Its key inputs are video streams from one or more cameras at each location, as well as repositories of generic or application-specific resources (for example images, graphics, pre-recorded video, or interactive components such as Flash movies). Repositories can exist on the server side to provide input of pre-recorded video streams for multiple clients. The layer's key output is the primary screen – in our case the living room TV on which video compositions will be rendered. The Audio Layer contains all the audio processing components. Its key inputs are audio streams from multiple microphones at each client, and its key outputs are speakers through which sound is reproduced – usually in a domestic living room. Finally, the analysis components take both audio and video streams to generate cues relating to the activity that is taking place at each client location – and hence sit between the Video and Audio Layers.

The following sub-sections focus on the key data capture, encoding and composition features of the system and explain how they provide functionality beyond the state of the art, moving from personal software focused on single users (such as Skype and Google+ Hangouts) to a system which is shared by groups at each location. Analysis and orchestration are beyond the scope of this paper, but further information about their functionality and performance can be found elsewhere [9][15].

4.1 Audio Communication

The goal of audio communication should be to enable the same user experience *as if speaking to someone in the same room*. This is especially challenging for communication between groups in a family living room, where clip-on microphones and headsets cannot be practically used, and the environment may be noisy and reverberant. The quality of phone calls today is still generally limited to mono at less than 4 kHz audio bandwidth, but user experience evaluations between groups have shown the importance of audio quality for the effective communication.

We obtained the results shown in Figure 5 from video conferencing tests in which two different audio systems were compared during social interaction between groups of people. Each group was exposed to "high" and "low" audio quality for a duration of about 5 minutes each and then asked to express their agreement to the statement *"The communication was natural and without problems"* on a 5-point rating scale ranging from 1 (strongly agree) to 5 (strongly disagree). The difference in subjective evaluation between high and low audio quality is striking.

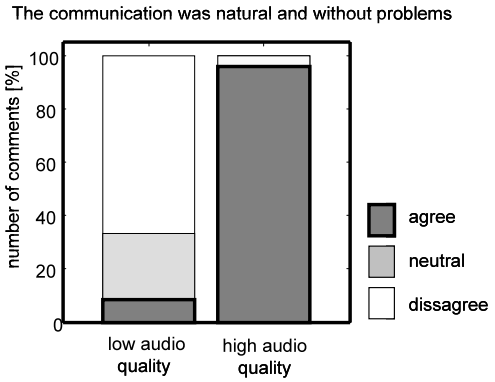


Figure 5: User experience during interactive group activity for low and high audio quality

The "high" audio quality used in the above tests - and in our Audio Layer - is a super-wideband Voice over IP (VoIP) system providing multi-channel audio at low delay and full audio bandwidth (24 kHz). The Audio Communication Component captures audio from an array of between two and four high quality microphones. It encodes and transmits the audio signals as two or three distinct channels (left, right, and optionally centre) using the MPEG AAC Enhanced Low Delay (AAC-ELD) audio codec – the same codec also employed in Apple's Facetime application. It performs echo control on the audio it captures and renders, and also allows external audio signals to be mixed with real-time audio signals, via stereo analogue inputs on the sound card to provide, for example, sound effects from a shared application. For transmission over IP, it incorporates Error Concealment (EC) and sophisticated Jitter Buffer Management (JBM) to handle packet loss and delay variations. By adaptively changing the playout time using Time Scale Modification (TCM) it achieves an optimal trade-off between late-loss and buffering-delay.

4.2 Video Communication

To provide effective coverage of social communication between groups, a higher quality video experience must be provided. A higher video resolution is necessary to provide both peripheral awareness of other participants and, when appropriate, eye contact plus the ability to transmit and interpret gestures and body language. Low delay is particularly important for the coherence of conversations, especially if they involve multiple participants. The use of multiple cameras also provides the flexibility to capture different views of each location, again improving the ability for the system to ensure that the activities of a group are effectively conveyed on the remote screen. While some commercial video telepresence systems do offer high resolution, low delay and even multiple cameras, they are designed for controlled room environments and optimised private networks – neither of which can be assumed in a domestic context.

The Video Grabber/Encoder in our Video Layer is a high performance component designed to capture images from an HD video camera, encode them and transmit them to the remote location with minimal added delay (below 100ms for an ideal end-to-end system). The Video Grabber captures digital or analogue video from a single camera using a hardware capture card. The SDI (Serial Digital Interface) standard is the preferred form of camera output, although signals can also be captured from low-cost HD webcam devices over USB. The Video Encoder is an H.264 encoder which is optimised for low-delay video transmission by using a number of techniques from the H.264 standard, including: not using B frames (bi-predictive pictures), using Constant Bit Rate (CBR) transmission, and minimising buffering throughout the system. A separate Video Grabber/Encoder is required for each HD video camera.

4.3 Audiovisual Composition

As previously mentioned, a critical aspect of combining a shared activity with audiovisual communication is the presentation of the game's features and the effective display of interactions among participants. This composition must be dynamic because the content will need to change as the focus of the activity and participants' roles develop.

The Visual Composition component enables our system to dynamically and seamlessly compose visual images at each location. It provides decoding and rendering to screen of real-time video at the lowest possible delay. It is capable of compositing the real-time video streams with pre-recorded media from a local repository. It can also incorporate other forms of content, such as Flash, from applications running locally. This enables a wide variety of visual presentations to be configured. It supports composition effects such as alpha blending for increased flexibility in presentation design. This component receives control and composition instructions from an orchestration module (or a faithful assistant). Composition information is expressed using the Synchronized Media Integration Language (SMIL) [3], which has been extended to define how real-time and pre-recorded media can be composited spatially and temporally.

Video feeds into the Visual Composition component are initially set up using RTSP (Real Time Streaming Protocol). Because live video is treated just as any other media renderer there is no architectural limitation on the number of simultaneous incoming video streams: the available hardware is the only factor determining how many video feeds can be displayed simultaneously. Video streams can also be decoded without display, in a form of 'standby mode'. This offers the significant advantage that, if control instructions decide to switch the currently displayed video to a stream that is in standby mode, the switch can happen instantaneously because the new video data has already been received and decoded. Setting up a new stream using RTSP would delay the switching time by least a few hundred milliseconds, and possibly significantly longer in practice in the case of aggressively optimised H.264 streams.

5. EVALUATING THE TECHNOLOGY

Following implementation of the prototype system described in Section 4, we carried out experiments to evaluate specific aspects of the data capture, encoding and composition components. As explained in Section 3, low-delay audiovisual communication and the dynamic composition of different camera views with game content were key requirements. We therefore chose to focus our technology evaluation on these requirements, and to provide comparisons with the state of the art where possible.

We have taken care to incorporate all components (e.g. from camera through to screen, or 'glass-to-glass') in our evaluations. This means that we have measured true end-to-end delays, and not isolated algorithmic or networking delays. For this reason, the results presented here are not always directly comparable to data published by other sources.

5.1 Video Transmission Chain

We measured the user-perceived round-trip delay of the video transmission chain as follows. Our experimental system comprised one endpoint in Amsterdam, The Netherlands, and another in Ipswich, England, both connected via the public Internet. These systems were equipped with the Video Grabber/Encoder and Visual Composition components described in Section 4.

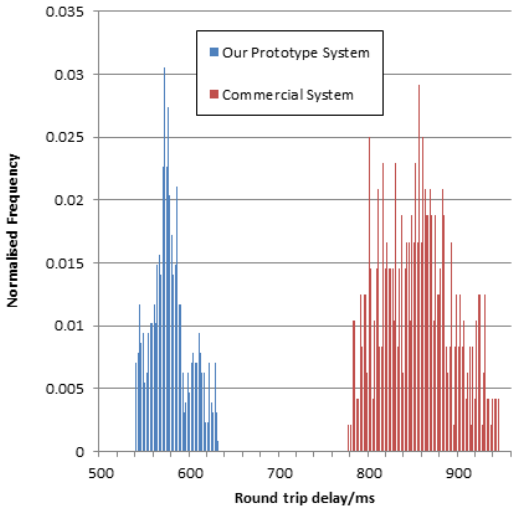


Figure 6: Round-trip video delay measurements

The network delay between the systems was 11ms. We took 500 measurements and measured an average delay of 580ms (with a standard deviation of 22ms); the distribution plot is presented in Figure 6. These are round-trip measurements including capture, encoding, transmission and display, and also the delays introduced by the measurement system itself. As the measurement system was measured to have a delay of 97ms (standard deviation 15ms) the actual end-to-end delay of our video chain is 242ms (standard deviation 37ms).

For comparison, we measured an equivalent commercial video conferencing system (using equipment from Polycom and Lifesize) connected through a conferencing bridge also via the Internet. This was the closest emulation of our prototype we could achieve using commercially-available equipment. The network delay between the systems was again 11ms. The measured round trip delay was 855ms (standard deviation 40ms); the distribution plot is provided for comparison in Figure 6. After correction for the measurement system delay, the actual end-to-end delay of the commercial system was 379ms (standard deviation 55ms). For this measurement we took care to use the similar codecs, bitrates and video sizes as we used for our own prototype.

While these results clearly show a significant improvement over a typical commercial system available today, the challenge remains to reduce the end-to-end delay as far as possible. In a laboratory environment, additional experiments were carried out to reduce delay times further, although these could not be incorporated in our prototype system.

A rolling shutter on the camera allows an image to be transmitted before a whole frame has been captured, and can reduce the delay by a fraction of the time it takes to capture one frame (25fps 40ms; 30fps 33ms, 60fps, 16.7ms). In addition, the Gradual Decoder Refresh (GDR) scheme enables I-frame information to be staggered over several frames, which reduces the bandwidth requirement and/or decreases the requirement for buffering and hence reduces the delay.

To effectively employ these techniques we developed an HD camera with a rolling shutter, and a capture card employing the above techniques and capable of handling the output from the rolling shutter camera. The glass-to-glass delay on the optimised video chain was measured at 85ms [13], a significant reduction on our prototype system.

5.2 Audio Transmission Chain

A different approach was required to perform measurements on the audio transmission chain, partly because the echo control in our prototype system meant that round-trip measurements could not be made. Instead, a PC-based oscilloscope was used to automatically measure the time difference between plots of a reference pulse which was passed into one Audio Communication Component, over a local network and out of a second such component. Several test runs of 1000 samples were carried out by repeating the reference pulse over long periods of time. The minimum delay end-to-end delay measured using this approach was about 52ms. It can be seen from Figure 7 that the distribution of delay values follows a repetitive sawtooth pattern. This is an interesting side-effect due to unsynchronised clocks at each Audio Communication Component. The resulting slight difference in sampling frequency produced an excess or deficit number of samples at the opposing sound card causing buffers to fill or empty respectively. The buffer control then dropped or created a frame when the latency grew too large.

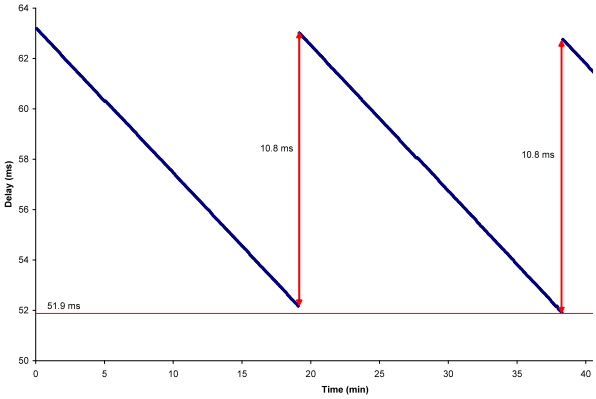


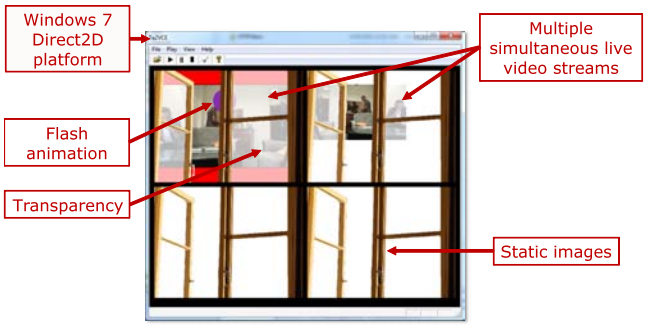
Figure 7: End-to-end audio delay measurements

If the measured network delay of 11ms for the Internet-connected systems used in the video delay tests is added to the delay observed here, the average end-to-end audio delay is approximately 70ms. While this could be reduced slightly by better clock synchronisation, it is still significantly smaller than the equivalent video delay. During subjective evaluation of our prototype system, it was necessary to artificially delay the audio transmission chain in order to achieve lip synchronisation between the audio and the video streams.

5.3 Audiovisual Composition

The Visual Composition Component is responsible for the seamless blending of visual streams, creating an immersive experience for the user where social communication and activities (e.g. gaming) become integrated. Because this functionality is very different to any other video communication system, no appropriate comparative performance metrics could be determined experimentally for the Visual Composition Component in isolation. Therefore, this subsection lists the component's features which were required to support effective integration of a shared activity, and which extend the implementation of visual composition which we have previously reported [13]:

- Aesthetic composition of real-time audiovisual streams and other pre-recorded media (text, graphics, video, Adobe Flash content)
- Both temporal (when to render) and spatial (where to render) composition
- Graphic overlays via an alpha channel with varying transparency
- Dynamic manipulation of visual elements, for example enabling external functions such as 'cut to camera'



(a) Composition of real-time audiovisual streams and other pre-recorded media

(b) Dynamic composition of visual elements during gameplay (e.g. cut to camera)

Figure 8: Visual Composition tests. The images above show key features of the Visual Composition Component, including the representation of different camera shots combined with game graphics.

Figure 8 shows some results obtained from tests of the Visual Composition Component, including the composition of real-time audiovisual streams and other pre-recorded media, and the dynamic composition of visual elements during game play (illustrating how camera sources can be switched within the same graphic composition).

6. EVALUATING SOCIAL INTERACTION

In addition to the technology evaluation described in Section 5, we carefully evaluated our system with real users in a representative social setting. This section describes the design of that evaluation and our findings.

Directly measuring togetherness is a difficult task: we wanted the users of our prototype system to be able to freely participate in social activities (without being wired with a dozen sensors), but we still wanted to obtain a number of objective results that would help us determine the added value of our technological choices. We felt that a comparison with existing conferencing technology was unfair (the use of better audio and video would highly skew the results in our favour, but say little about our approach in an abstract sense). We decided to compare interaction via our prototype with a more ground-truth-like togetherness experience: playing a board game in a physically co-located setting.

As discussed in Section 3, the basic interactive game model that we used was based on an outer space metaphor. Instead of developing one long game, we devised three short games, or *mini-games*. The mini games tested during the evaluation we: *Space Cruiser*, *Meteorite Girl* and *Pitch Matching*. The mini-game approach meant that a feeling of togetherness would be less dependent on the game logic: each mini-game was intentionally a short, focused and directed interactive experience. Collectively, we refer to this suite of mini-games as The Family Game.

The board game we selected was, like the Family Game suite, a turn-based cooperative game. The final goal was to evaluate how our system would compare to face-to-face game playing. We are aware of the differences between these two situations of game playing. The underlying game mechanics and in the interaction itself are different. The board game is played on a board, with playing cards and other game pieces and people communicate directly with each other (not-mediated), whereas Family Game is played through interaction with the Microsoft Kinect sensor and RFID playing cards - and people communicate through audio/video communication (mediated). Figure 9 illustrates some of the room infrastructure we used for running the Family Game evaluations.

6.1 Approach

The evaluation consisted of two parts: the evaluation of the *remote* Family Game, and the evaluation of the *co-located* board game. In order to perform the evaluations we used the Social User Experience Framework (SUX). This framework is intended for better understanding how people use and experience mediated social communication. The framework studies, among other issues, *interactics* (referring to people's experiences of interacting with the system and with others) and *aesthetics* (referring to the sensorial qualities of the system that enable social communication). *Interactic* experiences are based on human cognition (shortly) before, during, and (shortly) after interaction, and include the following constructs: *Quality of Communication*; *Social Connectedness*; *Challenge*; *Group Attraction*; *Inclusion of Other in Self*; *Overlap of Self, Ingroup and Outgroup*; and *Emotion*. *Aesthetics* are closely related to human perception, and include the following constructs: *Social Presence and Presence*,

RFID readers for interacting with the game (left), and main capture camera and Kinect sensor (right).

Figure 9: Room infrastructure for running the subjective evaluations

Naturalness, and *Immersion/Engagement*. Further information about this framework and the questionnaires used in our study are beyond the scope of this paper, but can be found elsewhere [20].

During the evaluation we used multiple methods for data collection, including several questionnaires:

- A general questionnaire beforehand (asking for information about the relationship of a specific person towards others in the group playing the game and to the group as a whole.
- Questionnaires after each condition (playing the Family Game or playing the board game).

Group interviews were used after playing both games to obtain qualitative information about people's experiences. Furthermore video recordings were made throughout the evaluations (in both conditions). All of our participants were paid a modest sum (€30) to take part in our experiments.

Each evaluation group consisted of four people who knew each other well (families or groups of friends). In total 36 people participated in the experiments. Half of the groups started with the Family Game (45 minutes) and then played the board game (45 minutes). The other half of the groups started with the board game and then played the Family Game. The youngest participant was 9 years old, the eldest 59. The mean age of the 36 participants was 22.5 years old. We realise that these 36 people represent only a modest sample group, but we feel that the results they provided are sufficiently instructive to warrant use.

Participants were invited to fill out several questionnaires:

- Characteristics of their personal relationship with the others and with the others as a group.
- Characteristics of their personal relationship as experienced at that moment (immediately after playing).
- *Aesthetic* experiences: Social Presence and Presence, Naturalness and Immersion/Engagement.
- *Interactic* experiences: Quality of Communication, Social Connectedness, Challenge and Emotion: affect (positive versus negative), arousal (relaxed versus aroused), and dominance (being in control versus being controlled).

The sessions closed with group interviews in which the focus was on how people experienced both conditions in terms of social connectedness and a feeling of togetherness.

6.2 Findings

Cronbach's Alpha [6] is widely used as a statistic to estimate the reliability of results from a sequence of experiments. Table 1 summarises the Cronbach's Alphas for the *aesthetic* constructs: SPP, N, and I/E. From this table we can conclude that the Cronbach's Alphas are consistent enough over different measures/evaluations. In reading the tables, it can be generally assumed that similar scores for the board game and the online, interactive Family Game means that the 'togetherness experience' is reasonable similar. This means that the infrastructure and the remoteness does not degrade the feeling of togetherness.

Table 1: Cronbach's Alpha for SPP, N, and I/E

	Family Game	Board game
Social presence/Presence	.899	.640
Naturalness	.851	.877
Immersion/Engagement	.897	.635

The use of Cronbach's Alphas means that constructs show 'internal validity' across measures for the aesthetic constructs. The Social Presence and Presence (SPP) for the board game is somewhat different from the other Cronbach's Alphas. This is due to the fact that one item (measuring the feeling of presence in the virtual situation) was lacking. Immersion/engagement (I/E) in the board game has a lower Cronbach's Alpha as well. We assign this result to the fact that in real life situations, or in this case playing the board game, the item related to 'feeling part of the activity' is interpreted differently than it is in a 'virtual situation' as is the case in the other measures. When this item is left out of I/E, the Cronbach's Alpha goes up and is .660, a little higher.

Table 2: Cronbach's Alpha for QC, SC, Ch

	Family Game	Board game
Quality of Communication (QC)	.866	.861
Social Quality (SC) (2 items remaining)	.889	.824
Challenge (Ch)	.785	.851

When we compare the Cronbach's Alphas for the *interactic* experiences, we see that the Alphas are very similar across playing the board game and playing Family Game — see Table 2.

This indicates that these constructs have sufficient 'internal validity' over both tested situations.

When we look at the aspects playing a role in long term relationships and how these are experienced over time (e.g. personal effort, thinking about each other, sharing experiences, staying in touch, recognition, and group attraction) we see that only the Cronbach's Alphas for thinking about each other and for group attraction carry sufficient internal validity — see Table 3.

Table 3: Cronbach's Alpha for thinking about each other and group attraction

	Before start of experiment	After Family Game	After board game
Thinking about each other	.700	.545	.603
Group attraction	.858	.903	.851

The reason why these two constructs are interpreted in the same way when asked in the context of one's relationship or after playing the game (Family Game or board game) is that they ask about an experience on a specific moment in time (e.g. a contact moment in the past and what happens after that in terms of experiences). The other constructs carry items that more generally ask a participant about their experiences of a social relationship over time, not a specific moment therein. When asked in the context of your relationship to the others, the items are therefore differently interpreted than when asked about your experience of a specific contact moment.

Table 4: Paired Sample Tests: Family Game versus Board Game

	Dif. Mean	S. Dev.	Sign.
Social presence/ Presence	-.714	.921	.000
Naturalness	-.676	.921	.000
Immersion/ Engagement	-.509	.845	.001
Challenge	-.396	.768	.004
Social Connectedness	-.287	.643	.011
Quality of Communication	-.456	.661	.000
Affect	.444	.909	.006
Arousal	.139	.990	.406*

If we look at the other constructs, we see differences in means between the Family Game and the board game, but most of these differences are not statistically significant — see Table 4. For most measures, the board game scores on average a little bit better. For the emotional arousal, the Family Game scores best and this difference is statistically significant. The emotional affect is also higher, but this difference is not statistically significant.

Related to the fun people had during both games, people often reported a preference for the board game, but they sometimes had more fun during playing the Family Game. We think this is largely related to the fact that the Family Game was more thrilling, and there was more tension in short periods of time than was the case during the board game. In the board game fun was often related to the social communication and a feeling of connectedness and direct contact. In the Family Game fun was often related to the thrill of playing the mini games.

The experience of the Family Game was more positive compared to the board game (though not statistically significant). Arousal was higher in the Family Game (statistically significant). Overall, it can be concluded that the Family Game did very well compared to the board game in these tests since there are only few significant differences. Though people tended to find a real life situation better and socially more enjoyable, the Family Game is a very good alternative to have high quality social interaction with others when no other means are available.

7. DISCUSSION

Current-generation video conferencing has provided a powerful communication tool for people and (to a more limited extent) families who wish to stay in touch while apart. In our work, we have looked at a next generation of domestic video conferencing, in which fluid interactions and higher quality audio and visual content - some of which is generated and composed dynamically - provide a sounder basis for a greater feeling of togetherness. Our work has consisted of doing a significant analysis of user requirements, followed by the development of a research prototype that has been deployed and tested in an international setting. We conducted a systems-level evaluation of the user interaction requirements for domestic multi-person, multi-party conferences. Finally, we conducted an analysis of the social aspects of perceived 'togetherness' compared with a baseline situation of intimate family interactions.

We can conclude that the original performance and functional requirements developed in Section 2 for domestic video communication were met. Our solutions advance not only the state of the art for home conferencing, but also provide advances over 'conventional' video conferencing in a business environment. In Section 5 we have shown advances in terms of audiovisual transmission and composition, while in Section 6 we have reported on a comparative study between of our system and a collocated board game. In general the results are encouraging, indicating that our system is a good alternative for family gatherings when apart.

During our work, we have gathered a number of lessons learned that we believe will be useful for other researchers attempting to follow our lead.

A key result, we feel, is that interaction in video conferencing requires more support than the efficient end-to-end transport of bits. Multiple information sources need to be gathered, selected and composed to support effective interaction. Our Visual Composition Component was designed to enable the separation of composition mechanisms and composition policies. This way, external components (manual or automatic) could be used to signal composition instructions, while reusing the core system. This allows for reusability in different contexts (e.g. supporting other activities) and conditions (e.g. reacting to other low-level cues).

Effective interaction will always be related to the underlying performance of the overall system. Calculating end-to-end delays is typically done at the network level. Unfortunately, this level alone does not measure perceived quality, since delays in processing and rendering the media are not considered. Typically, manual solutions (cameras and high-precision clocks) can be used for measuring, but these are tedious and prone to errors. In our research, we have determined the necessity to develop a standalone tool to obtain computable measurements of round trip delay times (glass-to-glass delay) for a complete end-to-end system. Further research in this area is still required.

The combination of the AAC-ELD audio codec with low-delay streaming and multi-channel echo control provided an extremely effective Audio Communication Component. The use of a 4-microphone array proved to be essential in providing directional audio. This support is essential if participants are to have the freedom of movement required to support interaction in an unencumbered manner.

While we are encouraged by the results of this work, it is clear that there are many nuances to supporting interaction that could be further developed. The context of the activity (a game, a birthday party, a shared remote dinner) could be used to influence the types of shots selected from the multiple cameras that are available in each location. This process could be manual (as in Villemard's vision in Figure 1), or it could be automatic. We suspect that some hybrid form will need to be developed first.

There is also scope to explore the applicability of our results for domestic multi-party video conferencing to very different contexts. One such example is the teaching of skills which require embodied learning, such as playing a musical instrument. The use of multiple cameras and dynamic composition could significantly improve the experience of a remote lesson conducted through a video conferencing system. We also anticipate that our approach could be applied to other domains including remote healthcare and semi-structured business environments.

Regardless of its type, not only will the context of the activity influence the selection of individual shots, it may also help influence graceful degradation of network connections between parties. More work is required to really understand how social togetherness can be maintained in the face of transient resource situations.

Finally, at the highest level, an in-home system should be responsive to the privacy needs of parties who are incidental to the shared interaction taking place. Unlike a managed office setting, the home often has multiple parallel activities taking place. These need to be protected and potentially exploited.

The advent of a new generation of super-fast broadband access networks, with increased speeds upstream as well as downstream, is making high quality domestic video conferencing viable in many countries for the first time. We feel that we have demonstrated the requirements and a solution for supporting increased 'togetherness' for domestic video conferencing – and shown that network speed alone is not enough to provide support for fine-grained home interaction. Our approach provides a valuable model for the development of future conferencing systems for both consumer and business applications.

8. ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of Orlando Verde and Wolfgang Van Raemdonck at Alcatel-Lucent, Rene Kaiser and his team at Joanneum Research and Dr. Marian Ursu and his team at Goldsmiths, University of London. This work was supported in part by funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760.

REFERENCES

[1] Ames, M. G., Go, J., Kaye, J., and Spasojevic, M. 2010. Making love in the network closet: the benefits and work of family videochat. In *Proceedings of ACM CSCW*, 145-154.

[2] Batcheller, A. L., Hilligoss, B., Nam, K., Rader, E., Rey-Babarro, M., and Zhou, X. 2007. Testing the technology: playing games with video conferencing. In *Proceedings of ACM CHI*, 849-852.

[3] Bulterman, D.C.A., Jansen, J., Cesar, P., et al. 2008. Synchronized Multimedia Integration Language (SMIL 3.0). *W3C Recommendation*. URL=http://www.w3.org/TR/SMIL/

[4] Chen, M. 2001. Design of a virtual auditorium. In *Proceedings of the ACM International Conference on Multimedia*, 19-28.

[5] Collins, R. 2005. *Interaction Ritual Chains*. Princeton University Press.

[6] Cortina. J. M. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98-1

[7] Dourish, P. 2001. *Where the action is: The Foundations of Embodied interaction*. Cambridge Massachusetts: MIT Press.

[8] Durkheim, E. 1971. *The elementary forms of the religious life*. Allen and Unwin.

[9] Falelakis M., Kaiser R., Weiss W., Ursu M.F. Reasoning for video-mediated group communication. 2011. *Proceedings of IEEE ICME*, pp 1526-1530.

[10] Gaver, W., Sellen, A., Heath, C., and Luff, P. 1993. One is not enough: multiple views in a media space. In *Proceedings of ACM CHI*, 335-341.

[11] Goffman, E. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Anchor Books.

[12] Groen, M., Ursu, M.F., Falelakis, M., Michalakopoulos, S., and Gasparis, E. 2012. Improving Video-Mediated Communication with Orchestration. *Journal of Computers in Human Behaviour*, 28(5): 1575–1579.

[13] Jansen, J., Cesar, P., Bulterman, D.C.A., Stevens, T., Kegel, I., and Issing, J. 2011. Enabling Composition-Based Video-Conferencing for the Home. *IEEE Transactions on Multimedia*, 13(5): 869-881.

[14] Kirk, D. S., Sellen, A., and Cao, X. 2010. Home video communication: mediating 'closeness'. In *Proceedings of ACM CSCW*, 135-144.

[15] Korchagin, D., Motlicek, P., Duffner, S. and Bourlard, H. 2011. Just-in-time multi- modal association and fusion from home entertainment. *Proceedings of IEEE ICME*.

[16] Lampi, F., Kopf, S., and Effelsberg, W. 2008. Automatic lecture recording. In Proceeding of the ACM international Conference on Multimedia, 1103-1104.

[17] Ljungstrand, P., and Björk S. 2008. Supporting group relationships in mediated domestic environments. *Proceedings of Mindtrek*, pp. 59-63.

[18] Nguyen, D.T. and Canny, J. 2009. More than face-to-face: empathy effects of video framing. In *Proceedings of ACM CHI*, 423-432.

[19] Ott, D.E., and Mayer-Patel, K. 2004. Coordinated multi-streaming for 3D tele-immersion. In *Proceedings of the ACM International Conference on Multimedia*, 596–603.

[20] Steen, M. (ed.). 2012. D8.8: User Evaluations of TA2 Concepts. *TA2 Project Public Deliverable*.

[21] Yang, Z., Wu, W., Nahrstedt, K., Kurillo, G., and Bajcsy, R. 2010. Enabling multi-party 3D tele-immersive environments with ViewCast. *ACM TOMCCAP*, 6(2): article number 7.

[22] Yarosh, S., Inkpen, K.M., and Brush, A.J.B. 2010. Video playdate: toward free play across distance. In *Proceeding of ACM CHI*, 1251-1260.

[23] Williams, D., Ursu, M.F., Meenowa, J., Cesar, P., Kegel, I., and Bergström, K. 2011. Video mediated social interaction between groups: System requirements and technology challenges. *Telematics and Informatics*, 28(4): 251-270.