Bringing Video Communication to the Community: Opportunities and Challenges

Ian Kegel
BT Research & Innovation, UK

Pablo Cesar
CWI: Centrum Wiskunde & Informatica, Netherlands

Marian F. Ursu
Goldsmiths, University of London, UK

Rene Kaiser
JOANNEUM RESEARCH, Austria

Jack Jansen
CWI: Centrum Wiskunde & Informatica, Netherlands

ian.c.kegel@bt.com, p.s.cesar@cwi.nl, m.ursu@gold.ac.uk, rene.kaiser@joanneum.at, jack.jansen@cwi.nl

ABSTRACT

The rise of online social networks, the wide availability of video communication technology and the deployment of high-speed broadband networks together provide the opportunity for video to become a medium for mass social communication among communities. However, current solutions provide poor support for ad hoc social interactions among multiple groups of participants. This position paper summarises the results of more than 5 years' research to make communication and engagement easier between groups of people separated in space. It shows how communication can be effectively combined with different shared activities, and how the technical capabilities of Communication Orchestration and Dynamic Composition work together to improve the quality of human interactions. The paper also describes ongoing work to develop the Service-Aware Network as a means of optimising the quality of a user's communication experience while making most efficient use of network resources. We believe these developments could enable video-mediated communication to become an effective and accepted enabler for social communication within community groups globally

INTRODUCTION

Video communication is not a new concept, and in its fifth decade it is only now gaining popularity with consumers, and predominantly in the form of two-way video chat between individuals. However, we believe that the convergence of several important trends in the near future will provide the opportunity for the growth of video as a significant medium for mass social communication among communities:

- **The rise of online communities:** The rapid rise of online social networking has displaced consumers from traditional instant messaging platforms. With synchronous communication on social networks increasing, applications such as Google+ Hangouts facilitate community-based group communication.
- **Ubiquitous low cost hardware supporting video communication:** All the technology components required to engineer a high quality video conferencing client are now available at very low cost in TVs, Set-Top Boxes, games consoles and standalone devices, and of course mobile devices too. Furthermore, Internet standards such as WebRTC[1] have lowered the barriers for application developers to leverage video communication.
- **The deployment of super-fast broadband:** Consumer broadband networks are still engineered for the delivery of more and more digital content from service provider to consumer, but increased upstream bandwidths arising from the roll-out of fibre mean that high quality video conferencing can be consistently achieved over domestic broadband.

Real world social communication will be complex and ad hoc in nature. Our collaborative research, both in the recent TA2 (Together Anywhere, Together Anytime) project[2] and the current Vconect (Video Communication for Networked Communities) project[3], proposes a set of capabilities that seek to go beyond today's video conferencing solutions to deliver dynamic, rich social communications with a great customer experience in the most cost-effective manner.

This paper introduces the key results of our research so far and the goals of our current collaboration, both of which we believe are highly relevant to the objective of this workshop.

RELATED WORK

Domestic video conferencing is becoming commonplace, with Skype providing a convincing existence proof of the viability of home video communication. Still, users encounter many limitations with existing technology. Recent studies 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

[1] www.w3.org/2011/04/webrtc
[2] www.ta2-project.eu
[3] www.vconect-project.eu

for multiparty viewing and use" [9]. These results corroborate our own research into user requirements [12].

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" [9]. Social games such as Mafia [2] provide another example, in which users value the ability to perform a shared activity together with remote parties.

One innovation is the presence of dynamic composition of audiovisual streams and content. This functionality has been identified in other works, highlighting the importance of manipulating and managing components within a set of video streams [5]. Studies on video-mediated free play between children found that different kinds of views led to different types of play [11], while other experiments demonstrate that good framing techniques improve social communication [10]. More recently, evaluations of remote game playing have shown that framing techniques can improve the effectiveness of the participants [6].

USE CASES

The TA2 project developed and evaluated several different use cases for group communication combined with a shared activity, based on a common set of technical capabilities.

One such use case, 'Family Game', investigated how a board game could be shared between multiple people at remote locations. Participants used the TV screen both for social communication and playing the game and their activity was captured by multiple microphones and cameras. Playing cards embedded with RFID tags were used to control chance aspects of the game, while the participants used their own bodies (via a Microsoft Kinect 3D motion sensor) as the interface for completing a series of activities. The game was intrinsically cooperative: players in different locations had to collaborate to achieve a common goal – for example collectively steering a ship through an asteroid field – by taking different individual roles which required communication. Figure 1 shows an example screenshot from the game.

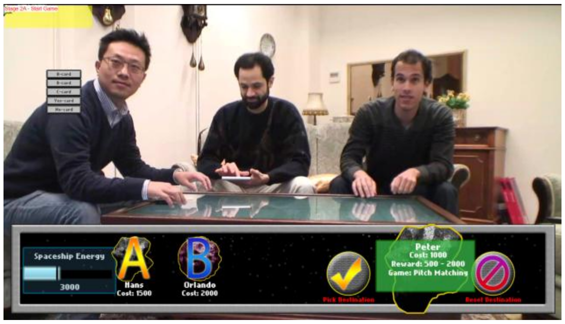

Figure 1: Screenshot from the Family Game prototype

While Family Game was built and evaluated in a laboratory environment, the TA2 project also developed a simpler single-camera prototype which was deployed for longitudinal studies in people's homes. This explored a different use case, 'Storytelling', in which family members were connected when physically distant. A grandparent could read a story book to a remote grandchild using two synchronised iPads. Pictures and sound effects from the book were simultaneously rendered on the TV screen over the video communication as a way of drawing the participants into each other's field of view.

Another important use case considered the teaching of skills which require embodied learning, such as playing a musical instrument. Institutions worldwide are beginning to use videoconferencing to bring scarce teaching talent to more and more pupils. The use of multiple cameras and dynamic composition has the potential to significantly improve the experience of a remote lesson conducted through a video conferencing system.

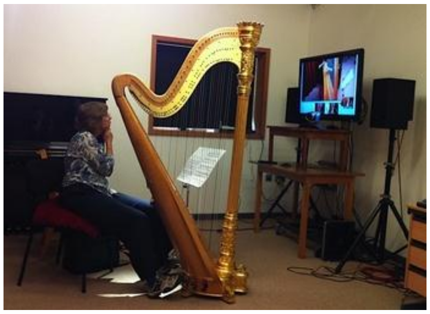

Figure 2: The Music Tuition prototype in action

COMMUNICATION WITH A SHARED ACTIVITY

Each of these use cases requires an infrastructure that supports high-quality interpersonal communication for consumer-oriented videoconferencing, sometimes involving multiple participants at each end. This section focuses on a key set of underlying technologies that allow for efficient media transmission, and for dynamic composition and manipulation of different media streams.

The infrastructure, according to our evaluations, made participants remotely playing a game feel as 'together' as collocated players [8]. For the evaluations, constructs of the Social User Experience (SUX) Framework were compared for both collocated and remote experiences. People found the collocated game more enjoyable, and there were no statistically significant differences in the level of 'togetherness' between collocated and remote experiences.

The system includes two main components, which in combination with Communication Orchestration (described in the next section) make our infrastructure unique:

- *Media Pipeline*: low-delay, high-quality audiovisual pipeline from grabbing to rendering that includes high-definition video and multi-channel audio;
- *Visual Composition*: a component that can dynamically and seamlessly combine the audiovisual

communication streams with content from a shared application or activity.

The *media pipeline* (data capture, encoding, decoding) included in our infrastructure enables a high quality video experience:

- It supports high video resolution, providing both peripheral awareness of other participants and, when appropriate, eye contact plus the ability to transmit and interpret gestures and body language.
- End-to-end delays are as low as possible. We succeeded in achieving an end-to-end average of 242ms (including visual composition), compared to an average of 379ms measured on a commercial video conferencing system [8].
- It allows for multiple cameras at each endpoint, providing flexibility for capturing different views.

All these capabilities ensure that the activities of a group are effectively conveyed on the remote screen. While some commercial telepresence systems do offer high resolution, low delay and even multiple cameras, they are designed for controlled environments and optimised private networks – neither of which can be assumed in a domestic context.

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. It integrates audiovisual communication with a shared activity by:

- 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’

COMMUNICATION ORCHESTRATION

Communication Orchestration can be described as intelligent camera selection behaviour which aims to improve social group communication and to support individual communication goals [6]. Following the metaphor of TV directing, orchestration refers to all the decisions that the director, cameramen, and editors take when composing a programme recounting a live event. As example, the shots illustrated in Figure 3 could be mixed to provide a more vivid and engaging representation of the activity in each of the two physical spaces.

Orchestration compiles a separate thread for each of the participating locations. Viewers are at the same time actors, and thus influence each other’s behaviour, and therefore what they should see and when they should see it.

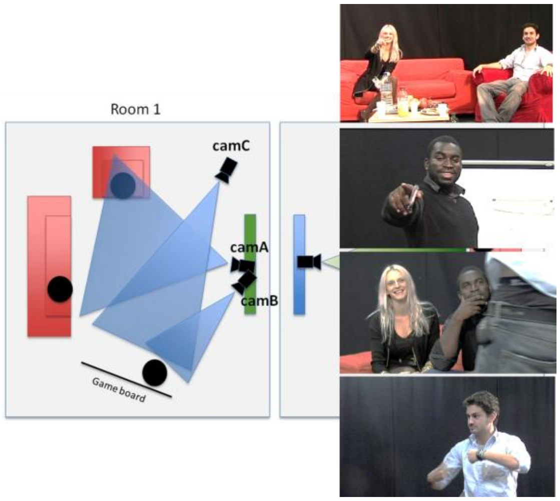

Figure 3: Functional shots that could improve the quality of a video-mediated interaction between groups of friends.

Orchestration decisions account for aspects of the *communication semantics*, including dynamic communication *structures*, the conversation *flow*, and the *social cues* used in communication management. Note that communication semantics does not refer to language understanding, but rather to the extraction of meaningful information *about* the conversation, such as recognising an attempt made by a user to utter a message, a turn shift, the pace of the conversation, the type of reaction (e.g. laughter) to something said or shown, etc.

In the TA2 project, we showed that Orchestration can improve the quality of the interaction between groups of friends socialising from different physical locations [6]. In the Vconect project we are exploring the same issue, but in larger and more complex communication structures.

Besides defining a body of Orchestration knowledge, i.e. a set of principles which can support communication in a certain setup, we developed a software framework for automating the reasoning process of communication understanding and decision making, taking a rule-based approach [13].

The Orchestration process outputs decisions regarding the use of the underlying communication infrastructure: i.e., the available audiovisual devices (cameras, microphones, capture and encoding processes, screens, speakers, and composition and rendering processes), specific transmission links, and video and audio multi-point control units. Therefore, it must always be aware of their capabilities, which may change in time, and of the means for controlling them.

THE SERVICE-AWARE NETWORK

Our research has shown that Communication Orchestration can substantially improve the effectiveness of social communication within a static group of people. Our ongoing work in the Vconect project is focused on a problem which naturally emerges from this: improving communication effectiveness in a similar way for users in larger communities whose needs are more complex. In these communities, subgroups may form and disband, participants' roles may change with respect to each other, and individuals may change their focus of attention rapidly. Another important capability, which we call the Service-Aware Network, is required to satisfy these dynamic communities.

The Service-Aware Network is responsible for implementation of decisions made by Orchestration in a manner which seeks to optimise the quality of experience for users in the 'real world'. Our vision is of Quality of Experience as a holistic concept to which technical limitations such as bandwidth and device capabilities contribute as much as the user's mood and emotional response to automated changes to the way video and audio is presented to them. The measurement and modelling of these parameters is a key challenge for Vconect.

However, we also believe that the Service-Aware Network has a complementary optimisation objective for the network or service provider. Today's video communication systems generally rely on static overlay networks that are either completely centralised, with all users communicating via a multi-point control unit or completely distributed, where users are fully interconnected with each other. These are suitable for a fixed communication model but cannot adapt themselves to the changing needs of a large, dynamic group. For example, consider the scenario of a 24-hour video chatroom service in which participants from anywhere in the world can join a multi-party video call at any time of day. Members of the chatroom may join and leave in an ad hoc manner and it could be expected that the geographical centre of their activity will move as daylight moves across the earth. The most efficient network topology to serve users connected at any one time will therefore change in a way which is not applicable to a single video chat session or a business videoconference. Another key challenge for Vconect is therefore to dynamically change the configuration of server components without interrupting the user experience – and more importantly to show that this could deliver a more cost-effective service among large communities.

In summary, we therefore believe that Communication Orchestration, Dynamic Visual Composition and the Service-Aware Network embody an important new direction for research in video-mediated communication which could enable it to move from a highly personal experience to become an effective and accepted enabler for social communication within community groups globally.

ACKNOWLEDGMENTS

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. Dourish, P. 2001. *Where the action is: The Foundations of Embodied interaction*. Cambridge Massachusetts: MIT Press.
4. Durkheim, E. 1971. *The elementary forms of the religious life*. Allen and Unwin.
5. 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.
6. 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.
7. 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.
8. Kegel, I., Cesar, P., Jansen, J., Bulterman, D.C.A., Stevens, T., Kort, J., and Färber, N., 2012. Enabling 'togetherness' in high-quality domestic video. In *Proceedings of the 20th ACM international conference on Multimedia (MM '12)*, 159-168.
9. Kirk, D. S., Sellen, A., and Cao, X. 2010. Home video communication: mediating 'closeness'. In *Proceedings of ACM CSCW*, 135-144.
10. Nguyen, D.T. and Canny, J. 2009. More than face-to-face: empathy effects of video framing. In *Proceedings of ACM CHI*, 423-432.
11. 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.
12. 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.
13. Kaiser, R., Weiss, W., Falelakis, M., Michalakopoulos, S., and Ursu, M.F. (2012). A Rule-Based Virtual Director Enhancing Group Communication. *ICME Workshops*, 187-192.

Ian Kegel is Head of Future Content Research at British Telecommunications plc. Having studied Electrical and Information Sciences at the University of Cambridge, Ian has worked in both the defence and telecommunications industries on projects ranging from radar signal processing to multimedia delivery, and has spent over 10 years leading research projects on digital media production and multimedia communication.

Pablo Cesar is a researcher at the Distributed and Interactive Systems group at Centrum Wiskunde & Informatica. He has (co)-authored over 50 articles about multimedia systems and infrastructures, social media sharing, interactive media, multimedia content modelling, and user interaction. He has given tutorials about multimedia systems in prestigious conferences such as ACM Multimedia, CHI, and the WWW conference (homepages.cwi.nl/~garcia).

Marian F. Ursu is a Reader in narrative interactive media and the Director of Research of the Department of Computing, Goldsmiths, University of London, UK. He has a first degree in Computer Science and Engineering from Technical University of Cluj-Napoca, Romania, and a PhD in Artificial Intelligence from Brunel University, UK. He leads the Narrative Interactive Media (NIM) research group at Goldsmiths (goldsmiths.nim.ac.uk).

Rene Kaiser is a key researcher at JOANNEUM RESEARCH, Graz, Austria. Rene studied Software Engineering at FH Hagenberg and is a PhD student at TU Graz. His main research interests are in the fields of telepresence and automatic non-linear video production, with particular focus on “Virtual Director” technology for automatic framing and selection of live video streams.

Jack Jansen is a researcher at Centrum Wiskunde & Informatica (CWI), with over 25 years of experience in multimedia and distributed systems. Empowering people to put available technology to a use they themselves envision is his driving principle. This results in activities ranging from languages, such as Python, via web standardization work (SMIL, Rich Web Application Backplane) to implementing systems for accessible and reusable multimedia (Ambulant). Recently, he has finally started to pursue a PhD (homepages.cwi.nl/~jack).