Smart Video Communication for Social Groups
The Vconect Project

Marian F. Ursu, Goldsmiths, University of London
Peter Stollenmayer, Eurescom GmbH
Doug Williams, BT
Pedro Torres, Portugal Telecom (SAPO)

Pablo Cesar, CWI
Niko Farber, Fraunhofer IIS
Erik Geelhoed, Falmouth University

***Abstract*—This article introduces the Vconect project. Vconect (Video Communications for Networked Communities) is a collaborative European research and development project dealing with high-quality enriched video as a medium for mass communication within social communities. The technical capabilities where Vconect innovates concern: high quality a/v capture, dynamic a/v composition, network resources optimization and communication orchestration. The project is driven by two main use cases. The first focuses on the integration of live video communication with social networking services. The second focuses on distributed performances, their automatic representation to remote spectators and the support for social interaction around such performances.**

Keywords—video conferencing, telepresence, social communication, group interaction mediated performance, active audiences

I. INTRODUCTION

Video communication is growing in popularity, with major manufacturers, such as Microsoft/Skype and Google, targeting the consumer end, group communication, and more comfortable modes of use. This is not at all surprising, as people naturally need to see and hear each other when they are in conversation. Body language and voice intonations are an essential part of the communication, in many instances more important than the words themselves. Furthermore, being able to show each other the objects and events about which we are talking, through video snapshots and audio recordings, significantly improves the quality of the conversation.

Social networks have taken the world by storm, having already become an intrinsic part of our social fabric. For example, in April 2014, Facebook reported 900 Million unique monthly visitors and Google+ 120 Million[1]. This is not at all surprising, as people are social beings: we need to form communities and interact with each other. Community interaction and belonging is an intrinsic part of our sheer existence and well-being. We have to share thoughts with each other, help, teach, play, or simply engage in idle chat.

Until recently, these two natural and complementary forms of mediated social interaction were more or less separated from each other. However, advancements have been made in integrating them, for example by Facebook, which has integrated Skype and is now offering point-to-point live video communication, and Google+, which is offering an integrated video group communication, via Hangouts. However, more is left to do for their full integration. As a matter of fact, "integration" is probably not the right term, as new forms of mediated social interaction could emerge by taking the two paradigms as starting points.

This paper presents Vconect, a project that focuses on the development of new forms of social interaction through live video, but also considering their integration with social networks.

Taking the perspective of live video communication only, Vconect considers two tightly-linked great challenges:

- provide for the *complex and dynamically changing topologies of social group communication (figure 1),*
- observe the *clients and network constraints with regards to audio and video capture, transmission, composition and rendering*.

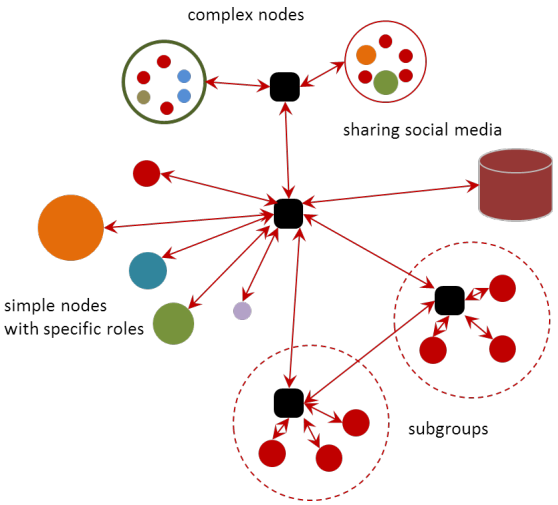

Figure 1: Example of a complex topology of social group communication

The former challenge refers to the ability of the overlay network to dynamically adapt, in aspects such as active

[1] http://www.ebizmba.com/articles/social-networking-websites

cameras and flexible modes of compositing and mixing the video content on each screen, to respond to aspects such as number of simultaneous users, the roles they have in the communication, the ability to separate into subgroups and have parallel conversations, the ability to deal with larger groups interacting with each other from different locations (see the figure 1 above for illustration).

The latter challenge refers to the ability of the overlay network to dynamically adapt in aspect such as encoding parameters (bit rates, resolution, formats, etc.) and transmission routes (to minimize delay) and to perform partial composition in the network, both view a view to optimizing the communication experience together with the cost of the network operation.

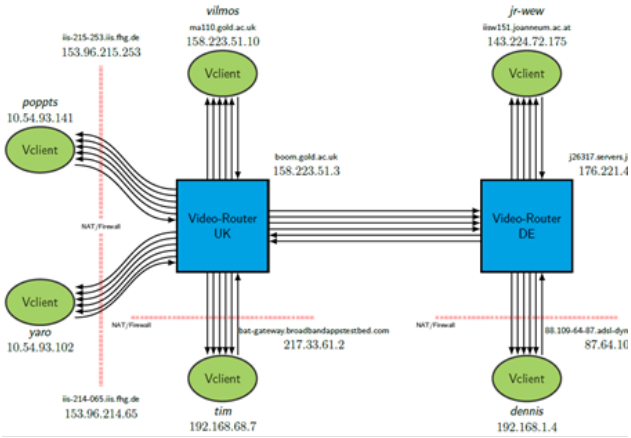

Figure 2: Example of deploying two video routers to optimise video traffic and minimise delay

There are various issues generated by each of these two challenges. However, they ought to be addressed concomitantly, if robust and effective technical solutions are to be developed.

The Vconect project[2] is an example of such an endeavor and the remainder of this paper will presented challenges as considered and solutions as devised by Vconect.

II. The Vconect Project

The Vconect vision is the adoption of high-quality enriched video as a medium for mass communication within communities.

Vconect is building a video communication platform which models and supports the complex communication topologies that characterize conversations between group members. The system takes intelligent decisions to mediate the communication at the level of audio-visual choices, screen layout and network capabilities. Vconect is ensuring the wide applicability of the platform by implementing, testing and evaluating it in the context of two different use cases. The first use case is based on the integration of video experience into social networking services, the second on group mediated performance.

[2] http://www.vconect-project.eu

Vconect's technological challenge is to develop components which enable a *service-aware* network. They must work together to intelligently and dynamically optimise network and media processing resources to satisfy the changing requirements of group conversations in communities. The requirements for high quality audio and video and low latency, which are inherent in a high quality experience, make this challenge even more demanding.

Vconect is advancing the state of the art in the following areas:

- Capture – Vconect allows for multiple cameras and multiple microphones at each end. This gives flexibility as to what is captured, including the ability to switch camera views, much like a TV director does when 'calling the shots', and also to choose what elements of the audio scene to capture.
- Composition – Composition relates to the way that the video and audio are presented to those involved in the group video chat. Vconect provides different views for different people and different interaction contexts; in some situations it may be good for all participants to see each other all the time; in other cases it may be best to see just one other person in full screen mode, for example.
- Network Optimisation – Vconect proposes the use of a *service-aware network* to facilitate transmission of audio and video streams. The service-aware network can dynamically decide where to place certain network components and should allow the service provider to control cost and the quality of experience for participants.
- Communication Orchestration – Orchestration decides which content is shown at each endpoint from the multitude of sources from the other endpoints, and also on how it is to be composited. The orchestration process has to be aware of the participants and the current communication context. It needs to know what audio video sources are available, i.e. what is being *captured*, it needs to collaborate with *network optimization* and it needs to instruct *composition*.

To help frame the requirements and test the innovations Vconect is working on two use cases:

1. The *Performance* use case – which links two sets of actors in two locations, through a video communications system in order to deliver one scripted performance. This is developed in collaboration with The Miracle theatre company in the UK.

2. The *Socialisation* use case – which adds group video communication capability to a social network designed for schools. This developed in association with SAPO/Portugal Telecom on their social network SAPO-Campus.

III. The Performance Use Case: Using Smart Video Communication to Support Innovative Multi-Site Theatrical Performances

Despite the increasing fidelity of recorded performance, be it of musicians, dancers or actors, live performances retain an enduring appeal. At the same time the nature of performance continually evolves, responding to the affordances provided by

technology. Two particular trends include that of live streaming performances from theatres to cinemas, so that stunning performances can be enjoyed by people unable to travel to, or be accommodated within, the venue housing the performance[3]. The second involves the audience moving between a number of performance spaces to find and appreciate different elements of the performance that they have to subsequently assemble to create their own stories from which they can derive meaning[4].

Vconect seeks to work at the intersection of these two; working with Cornwall based Miracle Theatre Company we are deploying our technology with a performance of the Shakespeare play The Tempest adapted such that the story is told through performances that take place in two separated venues, each with their local audience. Smart video communications technology from Vconect will provide: a means for the performers in the two locations to communicate with each other; for the actors in each location to be aware of both the local audience and the audience at the remote venue; and for the audiences in both locations to be aware of the performances in both the local venue and the remote venues and to allow a home based audience to enjoy a streamed version of the performance synthesised from the audio and video captured from the two performance spaces.

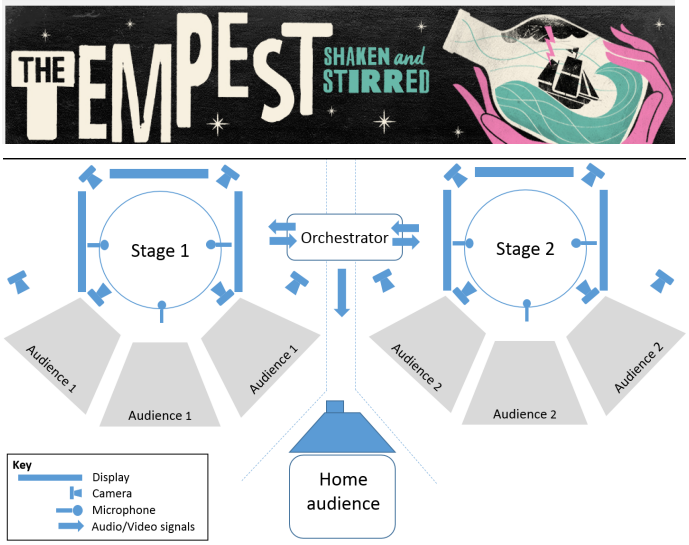

Figure 3: Schematic of the set up for The Two Site performance of the Tempest developed by Miracle theatre company showing that at each performance location multiple screens cameras and microphones are controlled using orchestration to control the way audio video signals are captured transmitted and displayed.

[3] Live Streaming of theatrical performances was pioneered by The Metropolitan opera in New York and has been embraced by amongst others, the UK's National Theatre (NT Live) http://ntlive.nationaltheatre.org.uk/ and by the Royal Opera House www.roh.org.uk

[4] Punch Drunk have "pioneered a game changing form of theatre in which roaming audiences experience epic storytelling inside sensory theatrical worlds". www.punchdrunk.com

Vconect technology is being deployed to support this use case that involves multiple cameras and multiple screens at each performance space. In addition new technology is being developed that will enable smart video communications to operate effectively in this complex scripted context. Tools are being developed to enable the writer to provide directorial instructions that can be translated into a machine readable instruction set that will control which audio video signals are selected for transmission, how they transmitted, and how they are displayed at each viewing location.

This deployment will allow Miracle Theatre Company to explore the challenges and opportunities associated with such a multi-site performance and will help them to pursue their goal of defining new genres of performance that mix theatre and film. At the same time it will help Vconect to develop tools that are well suited to the workflow required for theatrical performance.

The performances are scheduled for September 2014 and will take place in Cornwall in the UK.

IV. THE SOCIALISATION USE CASE - INTEGRATING SMART VIDEO COMMUNICATION INTO A SOCIAL NETWORKING SERVICES

The general use case targets the integration of real-time video communication with more asynchronous communication on social networks. It is formulated to understand how real-time video communication could enhance social networking forms of group communication, and to explore the dynamics of their relationship. For example, the way people migrate from one form to the other, how they (re)use resources between platforms, and what could be extracted from these interactions in order to improve the overall quality of the communication experience. Inherently, this case will explore more complex topologies of real time video communication than those supported by existing SoA systems, for example considering more communication nodes or subgroups, the latter referring to the ability of having a conversation within a smaller group whist still having presence in a/the larger social group.

This use case will be implemented through "SAPO Campus", SAPO being a brand of Portugal Telecom and Campus its platform for social media based learning, which targets schools and universities and can be used by both teachers and students. On top of the usual social networking features, like activity feeds of status updates, comments, forming groups, sending private messages, it includes SAPO's well-known services like blogs, photo, video and file sharing. Importantly the system also enables some of the functionalities of a learning platform, the ability to set homework, to submit homework and to keep track of a student's progress. The content and applications are fully owned and managed by the community with their own branding. SAPO Campus is different in nature from large well-known social networks as people in the same instance of the Campus are part of the same institution, providing a strong sense of community and responsibility. Moreover, it empowers schools as content providers giving them a public face and a single hub for their content.

A Vconect enabled video communication capability will be integrated with SAPO Campus, allowing seamless transfer between real-time video communication and the other currently supported forms of communication.

A key challenge for live video communication disclosed by this use case is the ability to support ad-hoc groups. Different groups, at different times, may initiate real-time video communication sessions between their members. To simplify the description, let us follow one of these groups only. Its members are video-connected, but at the same time they may have active links with other users on SAPO Campus. At the same time, the group may be visible as being engaged in a video communication. Other users, if allowed, may join the conversation. Existing members may leave. The main topic of conversation may branch off into a number of subthemes. They may generate the formation of subgroups, each subgroup being intensively engaged in conversations around its topic, but having the ability to "have a presence" in and maybe even "keep an eye on" on what is going on in the main group. As users can join and leave as they wish, this process leads to dynamic clusters that "travel" across the community, like swarms. Facilitating the communication needs in such dynamic structures is a challenge to address – this will be done under the heading of orchestration. Developing overlay media network optimisation techniques and the associated configurable media rich processes, necessarily required by such configuration, is another challenge – this will be done under the headings of service aware network and configurable a/v processes.

Actively connected members may decide to illustrate their points with sample media, which they access through the existing SAPO Campus interfaces. Combining the real-time video conferencing feature between groups with the ability to share time-based media is another challenge generated by this use case.

Finally, metadata extracted from the integrated communication platform (live video and SAPO campus) can inform its decision making process, such as the way recorded media is suggested or the way the real-time video communication is orchestrated.

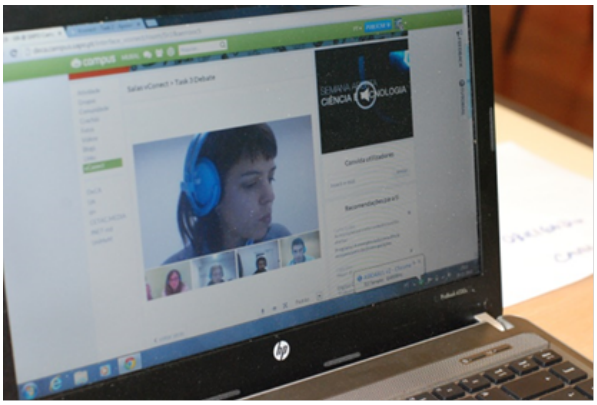

Figure 4: The Vconect video communications client integrated into the SAPO Campus social network site.

Through this summer (2014), Vconect is conducting a trial aiming at evaluating the integrated communication platform and the users' communication experiences it can support.

Vconect aims to provide an open web interface similar to WebRTC, which provides group video conferencing capabilities similar to Google Hangouts. The first successful implementation based on the Vconect API has been completed and the open API has been made available at Codebits 2014[5].

V. EXPERIMENTS TO UNDERSTAND THE MAIN REQUIREMENTS OF INTEGRATION OF VIDEO EXPERIENCE INTO NETWORKING SEVICES

Complementing the research driven by the two use cases described above, Vconect also explores the development of main technical capabilities through experimental enquiry. Three are summarised below.

A. View-modes and orchestration

A desktop screen could be organised such as to accommodate different video windows of different sizes in order to make the group conversation as easy and fluent as possible. Different such views will suit different people depending on the context. In some situations it may be good for all participants to see each other all the time; in other cases it may be best to see just one other person in full screen mode. As conversation roles change, it is very likely that a particular window should be used to show more than one group member, this leading to the requirement of mixing streams from different locations. Not only this, but the actual layout organisation may have to change in time, as the communication contexts change. We refer to the layout organisation of the live-video windows on the screen as *view-mode*. Orchestration is the process that decides how content from different sources should be mixed in each particular window, depending on the conversation context, as well as choosing a particular view-mode. An experiment was carried out that compared communication experiences in three main view-modes (see figure 5 below):

- Mosaic – participants could see everyone including themselves in a mosaic of video windows of equal size uniformly distributed on the screen.
- Unbalanced Mosaic with one main view plus a number of small windows below it (similar to Google+ Hangouts) – based on voice activity detection, the active speaker was displayed in a main window and the five remaining participants were displayed as a row of five tiles at the bottom of the screen.
- Full Screen – just one main window in full screen, in which, again based on voice activity detection like in the above condition, participants only saw the active speaker.

Figure 5: Candidate view-modes for group video communication

[5] https://codebits.eu/

A total of 54 volunteers were employed in this experiment, of which 18 were females (mean age 18.24, SD = 4.07) and 36 males (mean age 20.31, SD = 7.54). 16 were aged between 14 and 16 and 31 were aged between 17 and 30. There was a high predominance of participants that were under 20 years of age, as they were recruited from local secondary schools, colleges and universities.

Initial conclusions based on cluster analysis suggest that the Mosaic view mode is better suited to supporting fast turn taking, providing a sense of group cohesion whilst at the same time supporting the conveyance of the individuals' presence to each other (group presence). The Full Screen view mode appeared to be better suited to supporting slower-paced communication instances, of a more intimate nature, in which where there is mostly one person talking and an audience can see the facial expressions of the speaker in great detail. The Unbalanced Mosaic view mode is an interesting compromise between the other two more extreme cases, providing both for faster paced conversations and group presence, as well as slower paced conversations of a more intimate nature. Obviously, though, it also fails to provide best results in either of the two extreme cases. A more detailed description of the experiment and its finding is currently submitted for publication. Future experiments will also explore the effect of dynamically changing the view-modes through orchestration.

The three view-modes experimented with are currently supported also by Skype and the latter two by Hangouts. Nevertheless, on one hand, there is an open space of solutions when it comes to implementing the orchestrated behaviour of the system within these modes – i.e. how content is mixed within each window and how the view-modes change in real-time. On the other hand, there is also an open space for solutions regarding the way in which the network transmission is being optimised. In fact, it is the collaboration between these two reasoning processes that raise many research questions and provide space for innovation.

B. Social communication in living room setups

Video communication is not just about the desktop and Vconect is aiming to explore other communication setups such as, for example, sit-back communication via TV screens and multiple cameras that could cover living room spaces.

Vconect carried out an experiment based on a communication between three typical living room setups, each equipped with multiple cameras and a large TV set. Participants (6 per session) were invited to prioritise the qualities of their dream holiday and home in an informal context. They experienced two conditions:

- a split screen showing wide shots of the two other rooms and an orchestrated condition which provided.
- an mixed stream composed of wide shots and close up shots from the other two living rooms.

The shots in the second condition were automatically chosen according to a voice activity cue which was transformed into conversation turn taking information.

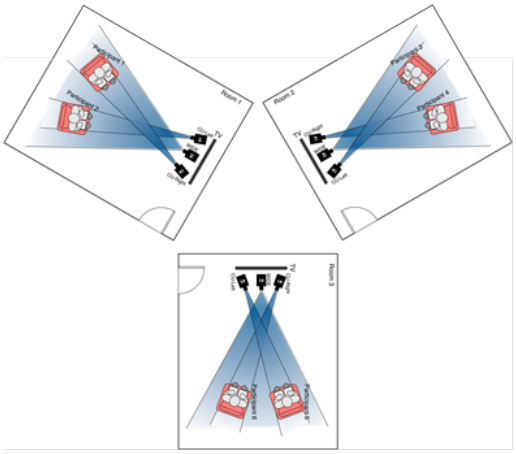

Figure 6: Room layout for the experiment looking into social communication in living room setups

A total of 24 volunteers took part in the experiment, of which 18 were female (mean age 22.89, SD 3.86) and 6 were male (mean age 22.83, SD 3.25) and they were all students at Goldsmiths, University of London.

The participants reported some very interesting effects of the difference between the orchestrated and static conditions. Many enjoyed the feeling of intimacy that emerged from seeing the detail of the close up shots but, at the same time, felt that a different segmentation of the communication space had occurred. For example, they considered that an intimate conversation was possible through the orchestrated condition, but not through the static split-screen condition. However, when the conversation was animated the split-screen condition was preferred as sometimes, when the rhythm of the communication was too fast, automatic mixing – that is orchestration – could not keep up with it. The more detailed description of the experiment and the corresponding results are currently being prepared for publication.

These insights combined with those gained from the View Mode experiments are suggesting that orchestration needs to optimise the need for seeing the active speakers as well as providing for group presence, but also for other, more subtle, aspects of the conversation. At the same time, orchestration should dynamically control the screen layouts (e.g. mix split screens with full screen).

C. Virtual microphone

The aim of this work is to capture the signals of a remote distant speaker using discreet arrays of static microphones. This is achieved through complex signal processing.

The functional design of the signal processing system capable of developing a "virtual" microphone using signals from an array of microphones has been built and tested in lab conditions. A recording of a man and a woman speaking together in the same room (not to each other just speaking at the same time) was played through loudspeakers. Algorithms to generate virtual microphones have been evaluated using a variety of internal parameter settings. The system assesses diffuseness and direction of the audio sound scene in time/frequency-space using the signals of the two arrays of

microphones as input. The signal processing then attempts to reconstitute an audio signal from a given location in the sound scene.

Figure 7: Photograph of lab set-up used for testing the virtual microphone performance

VI. CONCLUSIONS

Video communication for social groups is a large essentially unexplored domain. Initial solutions for simple setups have been built, particularly by Microsoft Skype and Google+ Hangouts, but these represent the early steps on this vast unexplored landscape. Vconect represents yet another step forward in this exploration. Vconect considers more complex communication topologies and is working on the development of a number of core technical capabilities, regarded as essential in supporting more complex communication structures, namely: encoding of high quality audio and video, dynamic audio-video composition, and automatic decision making processes able to adapt the communication infrastructure to the dynamic needs of social communication.

ACKNOWLEDGEMENT

Vconect is a collaborative European R&D project within the European Union's Seventh Framework Programme for Research and Technological Development. It receives funding from the European Community's Programme under grant agreement no. ICT-2011-287760.

We would like to thank all partners (in alphabetical order) – Alcatel-Lucent, BT, CWI, Eurescom, Falmouth University, Fraunhofer IIS, Goldsmiths, University of London, Joanneum Research Forschungsgesellschaft, and Portugal Telecom – for their inputs and comments.

Marian F. Ursu is Professor of Interactive Media at University of York, UK, and Professor of Computing at Goldsmiths, University of London. At York, he is the Head of Department Research. At Goldsmiths, he leads the Narrative Interactive Media research group. His research focuses on the development of new forms of video-mediated human-to-human and human-machine interaction, such as interactive narratives and smart video communication and telepresence.

Peter Stollenmayer is a programme manager at Eurescom responsible for the administrative and financial management of projects in the socio-economic and user related areas. This includes financial and administrative coordination of IST project NM2, a 3-year FP6 Integrated Project and TA2, a 4-year FP7 Integrating Project, and Vconect. He holds a M.Sc. degree in electronic engineering from the University of Stuttgart. Before he was with Eurescom, he worked for Deutsche Telekom in the area of ISPBX standardisation (1981-1999), and for the Engineering centre "Shape Technical Centre" in The Hague on next generation communications infrastructures (1989-1996). He was a member of the ETSI Board during 1996 and 1997. Peter Stollenmayer has had experience with management of European R&D projects for more than 15 years.

Doug Williams started his career in BT developing optical fibre amplifiers and switches. Since 2001 he has spent his time researching applications and services that may profitably occupy the data carrying capacity these fibres offer. Much of his work recent work has focused on TV related themes and has included projects on interactive narrative media; on using the TV to support group based games; on improving communication between groups and on calculating the aggregate bandwidth required when such new services are delivered to both consumers and businesses.

Pedro Torres is International R&D Coordinator at Portugal Telecom (PT), namely in SAPO Labs, one of the R&D branches of PT. He is involved in eliciting, planning and managing international R&D projects in partnership with research institutions and the industry in all the spectrum of information and communication technologies including areas such as information analysis, social media, video-conferencing, big data and web entrepreneurship. Pedro has managed PT's participation in projects such as Vconect (http://vconect-project.eu), LeanBigData (http://leanbigdata.eu) — on ultra scalable big data algorithms — and AppsForEurope (http://www.appsforeurope.eu) — on open data incubation and turning open data into viable businesses. Previously, he was a research fellow at Goldsmiths College, University of London, working for over two years on knowledge representation and automated reasoning in the FP7-funded TA2 project (http://www.ta2-project.eu) delivering video-conference to the home and, before that, he was a research assistant at Imperial College London in the areas of machine learning, automated reasoning and computation creativity.

Pablo Cesar leads the Distributed and Interactive Systems group at CWI (The National Research Institute for Mathematics and Computer Science in the Netherlands). 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.
http://homepages.cwi.nl/~garcia

Nikolaus Färber received his Doctoral degree in 2000 as a member of the Image Communication Group, University of Erlangen-Nuremberg, Germany. He has published numerous conference and journal papers in the area of robust video transmission and has contributed successfully to international standard bodies, such as MPEG, ITU, 3GPP, and DASH-IF. After being a Post-Doc at Stanford University in 2001 he joined Ericsson Eurolab, Nuremberg, Germany as a member of the speech processing group. Since 2003 he is with Fraunhofer IIS, Erlangen, Germany, where he is heading the Multimedia Applications department.

Erik Geelhoed is a Research Fellow at Falmouth University working on the Vconect program. Previously, with a background in psychology, he worked (1992 – 2009) at Hewlett-Packard Research Labs emphasising the importance of user requirements research in technology development and design. He worked with HP divisions, other large company's (e.g. Panasonic, Philips, Kodak), SME's as well as a number of art and community organisations during his time at HP. In Vconect Erik has conducted requirement studies for mediated performance with actors and dancers; audience research using Galvanic Skin Response sensors; user evaluations of automated editing (Orchestration) in social media settings; for Portuguese Telecom user research on social media. He is working with Cornwall based theatre companies to apply Vconect technology in performance settings.