Using Cinematic Techniques to Improve Video Communication

Rene Kaiser, Wolfgang Weiss

JOANNEUM RESEARCH
DIGITAL – Institute for Information and Communication Technologies
Graz, Austria

{rene.kaiser, wolfgang.weiss}@joanneum.at

Abstract

The Vconect project investigates novel ways of supporting mediated audio-visual communication for ad-hoc groups. This paper focuses on latest research on automating camera selection, a process called orchestration. The basic underlying problem it aims to solve is caused by the potentially high number of audio and video streams available in a group videoconferencing session. Each participant is typically covered by least one video camera, possibly more. Depending on the communication situation and the individual communication needs of the participants, a declaratively designed logic is executed that decides which video streams to show and how. In that realm, the Orchestration Engine executes cinematic principles automatically and within real-time constraints.

1 Introduction

In the following, we take a closer look at videoconferencing systems that aim to support ad-hoc social communication between groups of people – see Figure 1 for an example setup. One interesting problem implied by such video communication setups is that for each participant, there are multiple video streams available as options for being currently shown, i.e. when there are n participants and each is equipped with 1 camera, n–1 exterior video streams are candidates for being displayed at each client. The question is how to optimally deal with them. While one intuitive but not scalable option is to show them all at the same time in a fixed layout, we set out to investigate more sophisticated solutions with the aim of achieving better communication support through intelligent camera selection. Even though it can be argued that

audio plays an equally important role, the following will focus on video primarily.

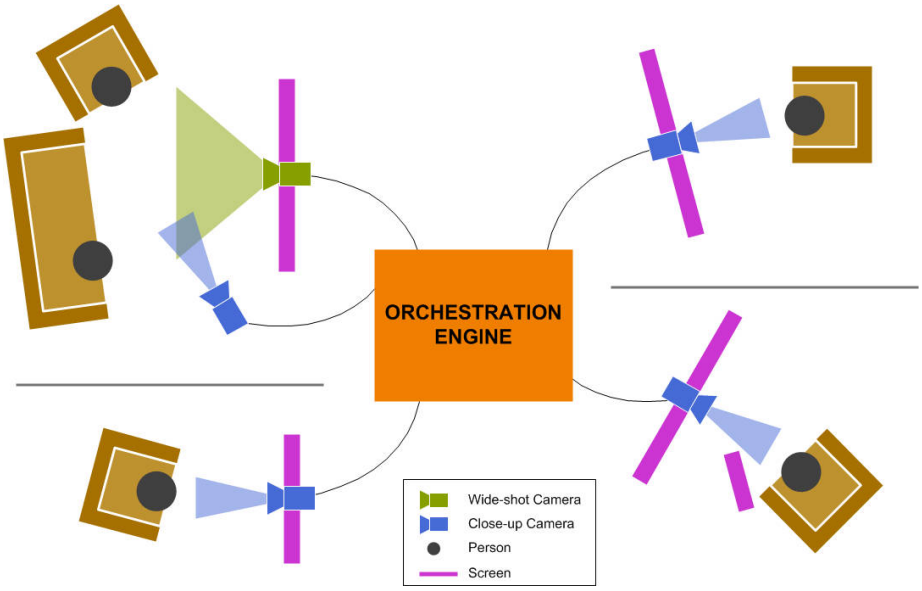

Figure 1 Schematic setup of a group videoconferencing session in bird-eye view. An Orchestration Engine as a central component observes and decides for all connected clients. Each location may involve multiple participants, cameras and screens.

Cinematic techniques – sometimes referred to as cinematographic techniques or principles – describe methods and common conventions used in video, film and TV productions. These techniques are the building blocks and creative methods which are applied by filmmakers to communicate meaning, to entertain, and to evoke a particular emotional or psychological response by the audience. This includes e.g. lighting, using depth of field, focus, camera position, camera movement, framing, special effects, shot cutting effects, etc. Subsequently, we discuss a set of cinematic techniques which can be implemented in an Orchestration Engine (cf. Ursu et al. 2013; Kaiser et al. 2012a), i.e. a software component deciding automatically which video streams to present how. A more general term also used in literature for automatic camera selection systems is *Virtual Director*, since it essentially aims at replacing a human director and camera operator crew. Selection of camera views is based on a set of principles that resemble those of the human professionals.

In the design process of this work, established cinematic techniques have been investigated, and some concepts have also been extended, for example

the selection of appropriate visual layouts (details in section 4). The system continuously needs to observe actions in the scene to choose the best camera views for the participants. Knowledge about the situation is available through low-level events as detected by real-time AV content analysis sensors.

A story can be told by cutting between one shot to another or to move the camera. Conveying the actions in a videoconference to each individual participant follows similar thinking. The Vconect setup does not involve movable (PTZ) cameras, but using high-resolution wide angle cameras allows to crop out interesting parts and to follow it by virtual panning, tilting or zooming. Cropped out parts of a physical camera view are also known as *virtual cameras*.

Different types of shots might be used in certain situations. Regarding its size, a shot type is defined by the distance between the camera and the subject. More concretely, it is the relation of the size of the visible part of the object to the total area in the shot. Relevant shot types for our Orchestration Engine implementation are close-up shots, medium shot and wide shot. If a certain shot type is not available e.g. a close-up of a person, it is also possible to use a crop of a wide shot for the desired person.

Note that while the participants' audio activity – who is speaking currently – is probably the most important information to reason with in such setups, also nonverbal communication cues are very important, especially mimics and gestures that can be interpreted as a reaction to somebody speaking. The following – possibly competing – high-level principles are guiding the design of orchestration behaviour:

- What's shown on the screen should reflect who is active, e.g. currently speaking.
- Awareness of the group, i.e. also of those currently not active verbally, has to be maintained over time.
- Reactions to what has been said are important ("active listeners").
- Self-awareness – knowing how oneself is seen and heard by others, and when.

Whenever we speak to someone directly, seeing and hearing the other person clearly is of utmost importance to support individual communication goals and to make the conversation a joyful experience. A high-quality full-screen close-up view might be shown in that situation, dimming or blocking other signals temporarily, enabling to grasp the other person's verbal and nonverbal reactions, like body-language, gaze and mimics.

However, automatically supporting that is a multifaceted research question. One aspect is to dynamically infer the communication situation, which

can be addressed by utilizing audio and video content analysis. A good understanding of communication theory is necessary to further understand how to support human communication by intelligent audio and video presentation decisions.

In this paper, we describe how users in a video communication can be supported by applying cinematic techniques. On that end, we aim to learn from motion picture directing and broadcast TV production. While it became clear in research experimentation (cf. Groen et al. 2012) that automatic editing for human mediated communication is a considerably different problem than automating for film production, nevertheless we can learn from film production grammar, and take out and realize individual principles to the benefit of the communication goals of the participants.

This work is embedded in the Vconect[1] project aiming to support complex communication topologies that characterise conversations between group members. Vconect is a successor to the TA2[2] project which also looked at supporting long-time relationships between geographically distant people via an audio-visual communication link – see Weiss et al. (2011) for details. Vconect focuses on two application domains: *Socialisation* aims to support informal conversations between students, embedded with the SAPO Campus[3] platform, while *Performance* investigates both mediated communication between artists conducting a performance in distributed fashion, and to the audience. Our technical approach is aligned with previous work on interactive live event broadcast in the FascinatE[4] project (cf. Kaiser/Weiss/Kinast 2012; Niamut et al. 2013).

This paper is structured as follows. The following section 2 gives an overview over several aspects that can be influenced by an Orchestration Engine through means of cinematic principles in a wider sense. Section 3 will take a closer look at the implementation of orchestration software, and section 4 presents an example in the realm of template selection. Conclusions, current limitations, and outlook to future work are discussed in the final section 5.

1 http://www.vconect-project.eu/

2 http://ta2-project.eu/

3 http://campus.sapo.pt/

4 http://www.fascinate-project.eu/

2 Orchestration Aspects

The following lists and discusses several detailed aspects that influence the decisions of an orchestration process in setups as described above. While some are not directly in the scope of cinematic techniques, they all have interdependencies and need to be considered when engineering orchestration behaviour.

Visual Presentation

Most directly, cinematic principles help to define what is visible on the participants' screens at each time. In each location, there might be one or multiple screens, personal or shared with others in the same room, and screens suited for specific purposes such as "second screens" (cf. Courtois/D'heer 2012). Screens might be added or removed any time, requiring the Orchestration Engine to adapt.

Templates, Screen Composition

A specific aspect of visual presentation is logic for dynamic layout selection and screen composition. The latter involves where on the screen to put video streams and other content and user interface elements. Using a limited set of predefined templates for their layout implies the advantage that users can get familiar with them and hence may more easily decide where to focus on. While well-known cinematic rules mostly refer to full-screen film editing, a lot of them can be executed in screens composited of multiple views as well. Further details on template selection will follow in section 4.

Dynamic Communication Topologies

Videoconferencing sessions among groups might be very dynamic in setup: people may join and leave a session any time, the number of participants in a certain location can change (cf. face/person detection video analysis modules). Subgroups might want to leave a session to move to a private discussion, possibly re-joining the former session at a later stage, and possibly seeking to keep peripheral awareness of the former session while having this side-conversation – comparable to a real human group would behave in natural face to face communication. Support for such dynamic topologies requires reasoning support from the Orchestration Engine and has strong influence on what is shown on the screens at each point in time.

Audio Playout

Besides the visual presentation decisions, audio orchestration is an aspect with enormous potential to enhance the communication experience, given that intelligent behaviour can be achieved. Currently irrelevant audio sources might be dimmed in volume – which certainly requires a great level and robustness in detection of communication situations and patterns to begin with. Using stereo playout can contribute to achieving a mental model of the spatial arrangement of remote rooms. Even more possibilities arise when more powerful capture technology is used, e.g. virtual microphones.

Integration with a Certain Task

The video communication itself might not be the only social process taking place during the session. In parallel, participants might conduct a collaborative task, such as a game. This implies requirements to integrate other visual components or video streams with a certain priority into the screens' dynamic content presentation. A particularly challenging task is joint media consumption, which might require synced replay of time-based media.

Adaptation to Device Type and Screen Size

Another aspect influencing visual content presentation are the technical capabilities of the playout devices and the size (resolution, aspect ratio, distance to users) of the screen area available to orchestrate for.

Automatic vs. Semi-Automatic Process

An Orchestration Engine might execute its behaviour in fully automatic mode, but may also allow users to steer and influence its decisions. User interaction can be regarded as an additional dynamic cue. It can take the form of very direct ("*display x full screen*") or indirect (e.g. choosing a certain cinematic style) interaction.

Supporting Special Conversation Roles

A further aspect influencing the execution of cinematic rules in a videoconferencing session are special roles of and relationships between the participants. A session might for example involve a teacher and several students – roles with which specific communication needs can be associated, to be supported by orchestration. Knowledge about both static and dynamic communication roles may further be exploited to predict turn-taking behaviour, i.e. predicting who is going to speak or answer next.

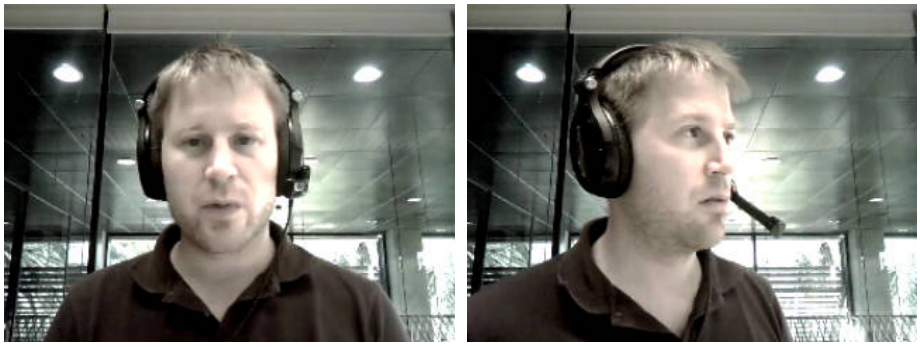

Figure 2 Two shots of a remote videoconference participant. The left one is a frontal shot that might be used when directly communicating with somebody. When available, the right one in contrast might be used when passively following a conversation of other participants, resembling the natural angle in a face to face group communication situation.

3 Orchestration Implementation

The Orchestration Engine in the Vconect project is a central, server-side software component reasoning for each participant to decide what they are going to see and hear from the available sources. It is continuously informed by events which are automatically extracted from the audio-visual streams or issued directly by the end users as requests or commands. Concrete examples are the audio activity of each person and when a user joins or leaves a session. These events allow us in a first step to continuously compute the state of the interaction and output higher-level events. In a second step, these higher level events are used to make actual decisions on what to show on each screen. The decisions of this process are forwarded to another component in the system that is executing the video routing and presentation.

The Orchestration Engine is implemented as a Web-based Java application running on the web server and servlet container Apache Tomcat. The application provides interfaces to external components to send and receive messages, such as from and to the user interface client via ActiveMQ. The logic of the Orchestration Engine is implemented as a set of rules which are executed by a hybrid reasoning system called JBoss Drools. This reasoning system executes both forward chaining and backward chaining rules, it implements interval-based time event semantics as described by Allen (1983) to allow temporal reasoning, and it is designed to process streams of events. In other words, JBoss Drools is an event processing engine (cf. Etzion/Nibbett 2010) and fulfils the requirement to make decisions in real-time.

Automatic orchestration can be seen as the process of intelligently selecting camera viewpoints based on low-level information. This process can be divided into two sub-processes realized by separate components in our implementation: the *Semantic Lifter* and the *Director*. The former processes low-level events to get an understanding of what is happening in the conversation. The latter is the decision making process which selects appropriate camera view points for each screen.

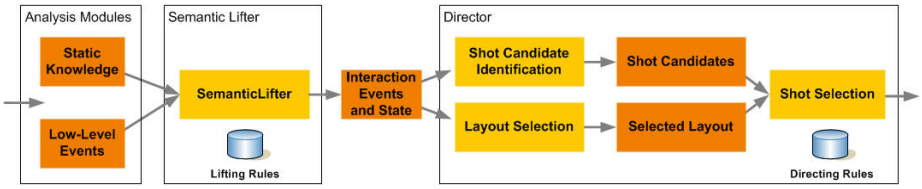

Figure 3 Architecture of Orchestration Engine, illustrating its information flow as a process from left to right.

3.1 Semantic Lifter

The input for the Sematic Lifter comes from various sources with different granularity and update frequency. For example, there are continuously generated events (cues) from the audio-visual analysis components from each location. The information available includes:

- audio activity events emitted by the audio analysis component at each client – by further processing this information, an understanding when someone starts and stops to speak can be achieved;
- position of the face, the eyes, nose and mouth within the video stream;
- information whether the eyes or the mouth are open or close;
- gender classification;
- recognition of facial expressions (“happy”, “surprised”, “angry” and “sad”).

The Sematic Lifter fuses the events from different sources in order to compute a global state of the interaction. Examples for the understanding of the communication state it aims to compute are “*Who is currently the most active person?*” or for identifying characteristics of the conversation “*How animated/heated is the discussion between the participants currently?*”. While managing internal facts and states, it lifts low-level cues to higher semantic level events based on definitions expressed declaratively as rules. The lifting step is necessary to bridge the semantic gap between the low-level sensors and the higher-level concepts to which decision making rules refer.

3.2 Director

The Director is the decision making component which processes the interaction events coming from the Semantic Lifter and selects an appropriate camera view point for each screen. This process can be further divided into the following subcomponents: Shot Candidate Identification, Layout Selection and Shot Selection. All decision making principles are expressed declaratively as rules as well in our implementation.

The Shot Candidate Identification creates and maintains a list of usable shot candidates based on the current conversational state, e.g. the currently speaking user might be the most interesting one, and therefore this component selects an available camera viewpoint for this user. Additionally, further virtual shots can be identified, e.g. a close-up view on the active speaker by cropping the face from the physical camera view. This component aims at fulfilling static aesthetical principles, e.g. "*Is the face within the camera view correctly framed?*" or "*Is the virtual shot moving smooth enough to ensure visual aesthetic expectations?*".

Based on current conversational characteristics, the Layout Selection will select an appropriate visual layout. For example, when there is a heated discussion of considerable duration where the participants talk successively in very short turns and a lot of crosstalk takes place, a tiled layout will be chosen which shows all participants in equal size, thus allowing seeing all other participants, which shall help to follow the overall conversation. In situations when a single person talks for a longer period, however, a layout can be chosen which gives the active speaker more attention, e.g. a full-screen view of that person.

The last step is the Shot Selection component which selects for each available space in the layout an appropriate camera stream. The aim of this component is to fulfil narrative aesthetic principles where appropriate sequences of camera views are selected. An example would be if two people are sitting next to each other and a cut from a close-up view one to the other is desired, a wide shot showing both of them might be used in between. The intended function of the wide shot would be to allow the remote participants to frame a mental model of the physical setup in that space (who is left and right), which ultimately is assumed to lead to a better communication experience.

4 Layout Selection Example

One concrete decision the Orchestration Engine continuously takes is the selection of a basic visual template. The following assumes video streams depicting participants only (see Figure 2) and disregards the integration of shared media etc.

Depending on the number of participants, the type of conversation, the current state and the current characteristics of the conversation, the Orchestration Engine will dynamically choose between several layouts. Conversation types can be divided into an intimate discussion, informal discussion or business meeting. Different states are applicable for certain conversation types, e.g. a business meeting can have an introduction of the participants, discussing the agenda or closing the meeting. Characteristics define how the conversation is going on e.g. if it is calm or heated discussion. The following layouts are available:

- The full screen layout, see also Figure 4a, might e.g. be appropriate when there is a monologue of a person talking to a large number of participants, or when a teacher speaks to his students. Initial experiments also revealed that this layout, i.e. sequences of full-screen cuts, are good for very intimate conversations.
- Figure 4b illustrates a full screen layout containing another view which is also known a picture-in-picture. This layout might e.g. be suitable when a private discussion between two people takes place. The large view is for the remote view while the small frame can be used for the self-view.
- A tiled view layout, see also Figure 4c, is suitable when there is a heated discussion where several people talk successively in a short time. In such situations where the conversation pace and the turn taking behaviour is very fast, full screen cutting is assumed to be too stressful to follow.
- Figure 4d illustrates a layout with one large view and several small views, similar to the one used by Google Hangouts. The large box can be used for the active speaker and the small tiles for all other participants.
- A layout with two larger views (Figure 4e,) and several smaller views might be suitable in a situation when there are two people intensively discussing and are alternating turns. The other participants might find this template useful to follow the discussion passively.

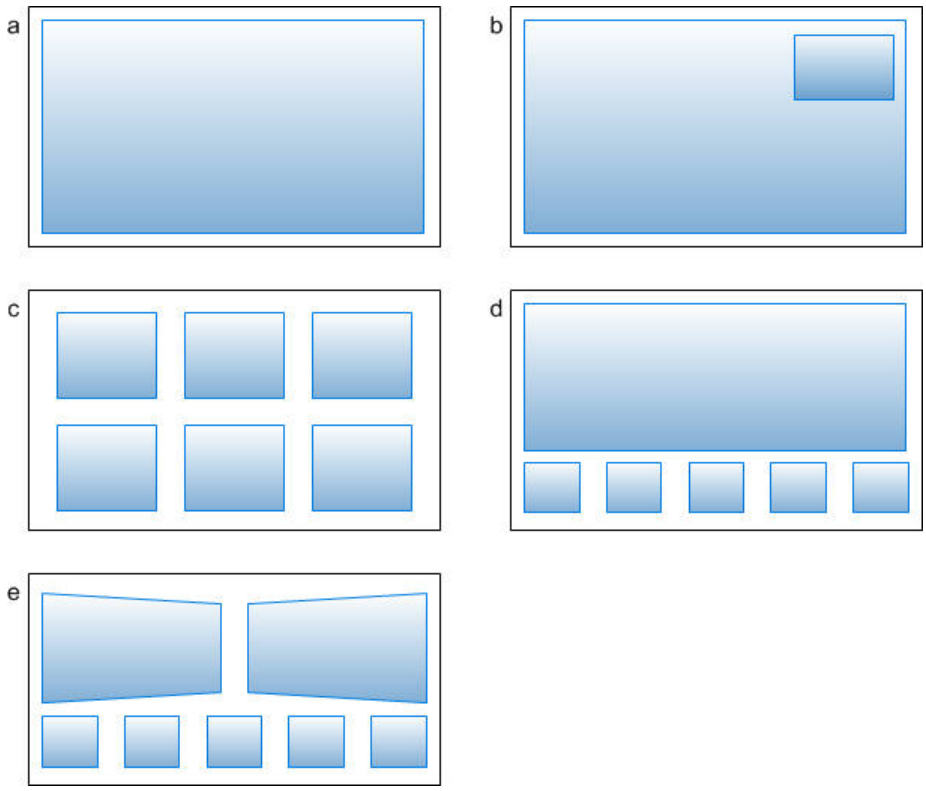

Figure 4 Possible layouts: (a) full screen layout; (b) full screen layout with a small tile as overlay of the main view; (c) tiled view layout; (d) focus on one person with small tiles for all other participants; (e) focus on two persons with small tiles for all other participants.

The possible situations mentioned above as suitable for certain layouts are mostly early hypothesis and by no means all verified by proper scientific experiments, however, this is a process the Vconect project is currently undertaking in order to build a rich conceptual framework of the orchestration space. Understanding which factors influence optimal orchestration decisions and how orchestration behaviour needs to be designed in order to lead to better communication experiences is a comprehensive research domain which needs to be investigated step by step.

5 Conclusion and Outlook

This paper presented on-going work on automating dynamic content selection for video communication. Research challenges and different aspects

influencing the application of cinematic principles have been explained. Our technical approach is based on event processing and rule-engines. The research prototypes implemented so far are naturally limited in scope and quality. The concept of human communication is very complex, therefore we seek to support group communication in certain setups involving certain tasks, thus reducing the problem space, and allowing focusing on orchestration behaviour for specific situations. By understanding communication situations and patterns, and hence choosing the right video streams to be rendered in certain layouts, the communication goals of the participants should be supported, and the communication experience should be enhanced. A useful concept in the evaluation thereof is *Quality of Experience* (QoE) (cf. Le Callet/Möller/Perkis 2013).

One practical limitation is the lack of cues. Integrating e.g. audio and video analysis components into such a system can be an elaborate task, and the quality of their results may not be as high as desired. We currently work on robust turn taking detection ("*Who is currently speaking?*") based on a rather simple audio cue (volume), which is quite difficult since the basic cue itself is very susceptible for background noise and change of microphone position. Another specific current limitation is the lack of nonverbal analysis cues. Integrating more analysis on this end would help to create much more detailed models about the conversation. Nonverbal analysis of social interaction is an ongoing research topic.

As immediate future work, we aim to focus on the following challenges:

- active use of ambiguities, vagueness, imprecision and uncertainty of information when processing it;
- prediction of turn taking, i.e. who is going to speak next;
- integration of the video communication with a social network, analysis of participant profiles to inform orchestration for the sake of a better communication experience.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2011-287760.

References

Allen, J. (1983): Maintaining Knowledge about Temporal Intervals. In: *Communications of the ACM* 28 (11), 832–843. Online: http://web.cacs.louisiana.edu/~logan/521_f08/Doc/p832-allen.pdf <2013-12-08>.

Courtois, C. ; D'heer, E. (2012) : Second Screen Applications and Tablet Users: Constellation, Awareness, Experience, and Interest. In: *Proceedings of the 10th European Conference on Interactive TV and Video (EuroITV '12)*. New York, NY: ACM, pp. 153–156.

Etzion, O.; Niblett, P. (2010): *Event Processing in Action*. Stamford, CT: Manning Publications.

Groen, M.; Ursu, M.; Michalakopoulos, S.; Falelakis, M.; Gasparis, E. (2012): Improving video-mediated communication with orchestration. In: *Computers in Human Behavior* 28, 1575–1579.

Kaiser, R.; Weiss, W.; Kienast, G. (2012): The FascinatE Production Scripting Engine. In: K. Schoeffmann et al. (eds.): *Advances in Multimedia Modeling – 18th International Conference, MMM 2012* (Klagenfurt, Austria, January 4–6, 2012). Berlin/Heidelberg: Springer, pp. 682–692.

Kaiser, R.; Weiss, W.; Falelakis, M.; Michalakopoulos, S.; Ursu, M. F. (2012a): A Rule-Based Virtual Director Enhancing Group Communication. In: *2012 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (Melbourne, 9–13 July 2012). Piscataway, NJ: IEEE, pp. 187–192.

Kaiser, R.; Weiss, W.; Kienast, G.; Bailer, W.; Thaler, M.; Thallinger, G. (2012b): Automatic Camera Selection for Format Agnostic Live Event Broadcast Production. In: A. Frotschnig; H. Raffaseder (eds.): *Forum Medientechnik – Next Generation, New Ideas : Beiträge der Tagung 2012 an der Fachhochschule St. Pölten*. Glückstadt: Hülsbusch, pp. 69–82.

Le Callet, P.; Möller, S.; Perkis. A. (eds.) (2013): Qualinet White Paper on Definitions of Quality of Experience. Lausanne, Switzerland, V1.2, March 2013.

Niamut, O. A.; Kaiser, R.; Kienast, G.; Kochale, A.; Spille, J.; Schreer, O.; Hidalgo, J. R.; Macq, J. F.; Shirley, B. (2013): Towards a format-agnostic approach for production, delivery and rendering of immersive media. In: *Proceedings of the 4th Multimedia Systems Conference (MMSys '13)*. New York, NY: ACM, pp. 249–260.

Schreer, O.; Macq, J.; Niamut, O.; Ruiz-Hidalgo, J.; Shirley, B.; Thallinger, G.; Thomas, G. (eds.) (2014): *Media Production, Delivery, and Interaction for Platform Independent Systems: Format-Agnostic Media*. Chapter 6: Virtual Director Technology. Wiley. To be published.

Ursu, M.; Martin, G.; Falelakis, M.; Frantzis, M.; Zsombori, V.; Kaiser, R. (2013): Orchestration: TV-Like Mixing Grammars applied to Video-Communication

for Social Groups. In: *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, October 21–25, 2013). New York, NY: ACM, pp. 333–342.

Weiss, W.; Kaiser, R.; Falelakis, M.; Mayer, H. (2011): Virtueller Regisseur in audiovisueller Gruppen-zu-Gruppen-Kommunikation. In: A. Frotschnig; H. Raffaseder (eds.): *Forum Medientechnik – Next Generation, New Ideas : Beiträge der Tagungen 2010 und 2011 an der Fachhochschule St. Pölten*. Boizenburg: Hülsbusch, pp. 194–205.