The Case for Virtual Director Technology – Enabling Individual Immersive Media Experiences via Live Content Selection and Editing

Rene Kaiser
Institute for Information and Communication Technologies
JOANNEUM RESEARCH
Graz, Austria
rene.kaiser@joanneum.at

Manolis Falelakis
Electrical and Computer Engineering Dept.
Aristotle University of Thessaloniki
Thessaloniki, Greece
manf@issel.ee.auth.gr

Wolfgang Weiss
Institute for Information and Communication Technologies
JOANNEUM RESEARCH
Graz, Austria
wolfgang.weiss@joanneum.at

Marian F. Ursu
Department of Theatre, Film and Television
University of York
York, United Kingdom
marian.ursu@york.ac.uk

ABSTRACT

An emergence of applications based on live audio-visual content streams could be observed in recent years. While the technological infrastructure in terms of bandwidth and device capabilities has advanced, media formats and related consumption paradigms have not changed as fundamentally. Meanwhile, a considerable amount of research has addressed automatic personalization of multimedia content for the sake of enabling immersive multimedia experiences, however, mostly considering pre-recorded and not live content. This paper states the case for more research to be conducted on what we refer to as *Virtual Director* technology as one key enabling technology for the hyper-personalization of live content delivery. A Virtual Director is software that automatically selects, frames, mixes and cuts from a number of AV content streams. It aims to automate the complex and challenging tasks that a broadcast director and team undertake during a live event broadcast. Virtual Director software can be applied in a range of use-cases, taking the individual's needs into account. There is unlimited scope regarding which factors such components could reason about in decision making. While such Virtual Director software has been developed as research prototypes, manifold challenges remain open to unlock its full potential. This paper presents recent technological achievements and reflects the potential of the approach in two selected application domains, interactive live event broadcast and group videoconferencing.

4th International Workshop on Interactive Content Consumption at TVX'16, June 22, 2016, Chicago, IL, USA.

.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications—*Computer conferencing, teleconferencing, and videoconferencing*

Keywords

Virtual Director; social multimedia; telepresence; cinematographic principles; live event broadcast; camera selection; viewpoint selection;

1. INTRODUCTION

The increase of live video stream services is very visible to consumers through a high rate of new multimedia applications with ever-improving audio-visual quality. The available bandwidth allows transmitting high resolution streams with low enough delay. Still, most services follow a broadcast model and do not aim to deliver a truely personal experience by taking individual user preferences into account. Relatively few commercial systems are adapting content on an atomic level, aiming for hyper-personalization. The value of such capabilities, however, is unquestioned.

Research prototypes have been developed that aim to address this problem space. *Virtual Director* software metaphorically aims to mimic and automate the work and knowledge of a live TV broadcast team. This concept [5] [14] [1] is a key enabler for immersive experiences on top of multimedia systems. Beyond basic tasks of automatically selecting, framing, mixing and cutting from a number of available live media streams, existing multimedia experiences can be enhanced through new levels of personalization, automatic content adaptation to playout devices, etc. Such components are required to take decisions within real-time constraints to deliver more interactive media consumption forms.

The real-time aspect is key since what's happening in a scene observed by cameras can't be predicted for the most

part. User preferences may change dynamically as well, limiting the scope to compute options in advance. Via services realized with a Virtual Director approach, every user may get different content, and user profiles or preferences expressed through whatever (abstract) form of interaction with the system could be changed during consumption as well, to which the systems needs to respond immediately.

Virtual Director technology can take decisions for an individual user, however, it can also enhance experiences for social groups, in a co-located space or distributed in several places. Example scenarios range from rather passive in nature, e.g. watching remote theatre performances together with friends, to rather active, as in a group of friends attending a language course via videoconferencing. A Virtual Director can either decide on media presentation for social groups as a collective, not just for each node individually. It can help create an immersive social experience in which geographically distant people feel part of a group, or combine any activity with a social communication link.

On a technical level, across multiple application domains, we aim to build a generic Virtual Director software framework, using a rule-based approach with event processing technology. Two main challenging aspects on the technical level are (i) to make use of low-level sensor information in order to achieve an *understanding* of the scene that is covered by the media streams, and (ii) to execute a set of pragmatic and cinematographic principles for *decision making* in real-time.

As an example in video-mediated communication between larger groups, low-level cues from speech (audio) and face detection (video) are interpreted to understand communication patterns in a process called *Semantic Lifting* (first challenge). Based on that, the Virtual Director decides what to show to each individual participant, taking cinematographic rules into account in deciding when to cut to another camera (see [1] and references therein; second challenge). The behavior can ultimately lead to immersive experiences and contribute to effects like *telepresence*.

The following sections discuss two application examples, from the domains of videoconferencing and live event broadcast, to illustrate how automatic content selection on live audio and video streams can have an impact.

2. GROUP TELEPRESENCE

We have built a Virtual Director system (called Orchestration Engine) for social group communication in different setups, using multiple microphones, loudspeakers, cameras and screens. The design of our system for social group communication has been informed by higher values such as *togetherness* [12]. See Figure 1 for an example setup. Based on that system, a number of evaluation experiments have been conducted and published [3] [15] [14] [2].

To support telepresence and other communication goals in such a context, the quality of a number of factors is key to enable a conversation that is both effective and enjoyable [8]. Audio and video have to be in sync, delays greater than 200 milliseconds are generally regarded as disturbing — see [13]. Audio is regarded more important than video in group communication, however, a certain balance needs to be maintained regarding the different modalities. Audio can be enhanced by using microphone arrays and quality enhancing features (e.g. dimming background noise, echo cancellation). The resolution of transmitted videos is important

Figure 1: Evaluation setup for social video communication with four people in two rooms that are equipped with one screen and three physical HD cameras each. The Virtual Director decides which video streams to render on screen depending on the communication situation.

for a number of reasons: if gestures, facial expressions, eye gaze, body language and such can be recognized, it makes the communication more natural, closer to face-to-face communication (cp. media naturalness theory [7]). Further, observations from our trials indicate that the ability to show real-life items to people in remote locations is crucial. The distance of participants to screens and the size of remote heads are also relevant to achieve a natural communication atmosphere.

A Virtual Director is needed in such systems simply because there are too many media streams to play in parallel. But beyond that pragmatic reason, such technology can enhance the experience, and also influence the social communication through its decisions. Popular solutions like Skype and Hangouts generally work well for a limited number of participants, but have clear limitations regarding camera selection behavior compared to natural face to face conversation or professionally edited motion picture. Further, unlike a real space where people meet and chat, the conversation topology in today's solutions is rather constant, has no direct way of branching into side-conversations without leaving the current communication space, to maintain lateral awareness of such.

We aim to continue to evaluate the effectiveness (*Are individual communication goals met?*) and immersiveness (*Do participants like the experience?*) of Virtual Director approaches in further experiments. One string of experimentation is to further look into the human capabilities that machines try to replicate in this context. Human directors naturally benefit from their implicit knowledge and understanding of the conversational context and a rich set of verbal and non-verbal cues. The Virtual Director is handicapped in this sense, having to work with the inherent limitations of real-time AV analysis or other sensors, in terms of number of features and detection accuracy, and the closed set of rules.

We have started to explore other sources that could inform the decision making process, for example extracting information from a social network that a videoconferencing environment could be integrated with (see [10]). The aim behind this research was to see how conversation patterns can be predicted and which factors influence them.

3. PERSONALIZED LIVE EVENT BROADCAST – 'NARROWCAST'

In a second application domain we have been working on a Virtual Director for live event broadcast, based on a scene-capture approach with a panoramic camera [11] and multiple microphones [9]. Note that the well established term *broadcast* is used here even though our aim is at a personalized *narrowcast*. While in traditional broadcast every user receives the same – except e.g. screen size and distance, color calibration, audio playout quality, aspect ratio cropping – in setups employing a Virtual Director component every user or group of users might receive a personalized output, e.g. including a bias to a sports team, person, or type of action. Personalization means taking user preferences into account along all subprocesses from sensor interpretation and camera framing to decision making regarding when to cut from which viewpoint to another. It also concerns the user's infrastructure, for example adapting to the screen size. The ideal zoom level and panning speed is very different when you intend to produce for either cinema projections or small mobile phone screens, yet few live productions take this into account and produce for multiple output channels in parallel.

A sample scene is depicted in Figure 2. The system automatically produces individual content streams for different playout devices and user preferences. Viewers may watch different parts of the scene according to their interests – while for some content types like sports there is at least a superficial common understanding what is most relevant and how to frame it in video, for performance shows like the one depicted there are few rules.

On one hand, several quality factors like content transmission delay play a lesser role in this domain compared to videoconferencing since media is sent in one direction only. In the other hand, we found in user evaluations with both production professionals and users without such particular knowledge that expectations regarding the visual cinematic quality are based on professionally edited TV programmes – some of which are fully scripted, so every camera movement and cutting decision can be planned.

The behavior of our Virtual Director prototype [6] for this domain was crafted with limited production grammar engineering resources. We concluded that its decision quality could not compare with the professional live editing skills and is especially lacking in perceived creativity, storytelling skills and intuition. Nevertheless, a Virtual Director provides the advantage of quicker reactions to low-level cues, which seems to play an important role in such a setup, and a consistency in decision-making that results in a more reliable experience. Due to factors such as fatigue, difficulty in hearing and seeing the events in the scene, and inherent differentiation in human mixing responses to salient events, human broadcast professionals will not be consistent in their decisions, while a machine ceteris paribus always responds the same.

We further conclude that purely reactive behavior has certain limitations and predictive situation/scene understanding is desired for future research iterations. We argue that the added value of our concept for enhancing multimedia experiences lies in the parallelization of individually tailored content selection decisions on a scalability level that a human production team can't realize for economic reasons.

4. DISCUSSION AND OUTLOOK

We have illustrated the potential of Virtual Director technology in the context of immersive multimedia applications using live audiovisual content streams. There are still lots of issues to be addressed by future research activities. For example, it is difficult to extract all the necessary information from low level cues or to structure a comprehensive set of cinematographic rules. Humans naturally benefit from their implicit knowledge and their feelings to foresee certain situations, which is very challenging for software components to replicate.

We have implemented Virtual Director research prototypes in two different application domains. The domain of group videoconferencing benefits from a Virtual Director through better *communication experience* by the participants. We hypothesize that the benefit is bigger the more complex the setup is, either regarding number of participants, or cameras/screens per participant, or any of the many other aspects in such setups. A Virtual Director in the domain of interactive live event broadcast enables mass customization in content production where viewers can individually select *what* and *how* to watch. Ongoing research efforts should lead in both application domains to *new forms of interactivity* and a more *immersive multimedia experience*.

Other application domains for Virtual Director approaches include specific group communication scenarios like refugee support, remote learning (e.g. in massive open online courses, MOOCs), telehealth/telemedicine, remote care, distributed theatre performances [4], and new forms of participative democracy. The approach appears to be especially relevant for novel types of media content, e.g. panoramic and 360 ° video content, or live content for virtual reality (VR) playout devices. In the VR domain especially there are issues that prevent users from watching lengthy content, i.e. virtual reality sickness (cybersickness, motion sickness). Even if these issues get solved, the challenge remains how to enable users to switch between *lean forward* interactive content consumption and *lean backward* passive watching. A Virtual Director approach might be very useful in such scenarios.

Overall, it can be stated that this research area is still in its infancy, but given its obvious potential, more research needs to be conducted to deliver components that can serve users in real scenarios outside research labs. On a more detailed level, remaining research challenges are for example to consider the scalability of the approach in applications that require distributed decision making, standardized representation formats for Virtual Director behaviour, tool support for the authoring of Virtual Director behaviour, and design patterns that enable the decoupling and re-use of bodies of Virtual Director behaviour.

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements no. 214793 TA2[1] – *Together Anywhere, Together Anytime*, no. 248138 FascinatE[2] – *Format-Agnostic SCript-based INterAcTive Experience*, no. 287760, Vconect[3] – *Video Communications for Networked*

[1] `http://www.ta2-project.eu/`
[2] `http://www.fascinate-project.eu/`
[3] `http://vconect-project.eu/`

Figure 2: Performance in a large stage as captured by a panoramic video camera system. The Virtual Director in this setup automatically framed a set of animated virtual cameras within the panorama and took different automatic cutting decisions for different users in parallel.

Communities and no. 610370, ICoSOLE[4] – *Immersive Coverage of Spatially Outspread Live Events*.

5. REFERENCES

[1] M. Falelakis, M. Groen, M. Frantzis, R. Kaiser, and M. Ursu. Automatic orchestration of video streams to enhance group communication. In *Proceedings of the 2012 ACM MM International Workshop on Socially-Aware Multimedia*, pages 25–30. ACM, 2012.

[2] M. Falelakis, M. F. Ursu, E. Geelhoed, R. Kaiser, and M. Frantzis. Connecting living rooms: An experiment in orchestrated social video communication. In *Proceedings of ACM TVX '16*, 2016.

[3] M. Groen, M. Ursu, S. Michalakopoulos, M. Falelakis, and E. Gasparis. Improving video-mediated communication with orchestration. *Computers in Human Behavior*, 28(5):1575 – 1579, 2012.

[4] R. Kaiser, M. F. Ursu, M. Falelakis, and A. Horti. Enabling distributed theatre performances through multi-camera telepresence: Capturing system behaviour in a script-based approach. In *Proceedings of the 3rd International Workshop on Immersive Media Experiences*, ImmersiveME '15, pages 21–26, New York, NY, USA, 2015. ACM.

[5] R. Kaiser and W. Weiss. *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media*, chapter Virtual Director. Wiley, 2014.

[6] R. Kaiser, W. Weiss, and G. Kienast. The FascinatE Production Scripting Engine. In *Advances in Multimedia Modeling*, volume 7131 of *Lecture Notes in Computer Science*, pages 682–692. Springer Berlin Heidelberg, 2012.

[7] N. Kock. The psychobiological model: Towards a new theory of computer-mediated communication based on darwinian evolution. *Organization Science*, 15(3):327–348, 2004.

[8] P. Ljungstrand and S. Björk. Supporting group relationships in mediated domestic environments. In *MindTrek '08: Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, pages 59–63, New York, NY, USA, 2008. ACM.

[9] O. A. Niamut, R. Kaiser, G. Kienast, A. Kochale, J. Spille, and O. Schreer. Towards a format-agnostic approach for production, delivery and rendering of immersive media. In *ACM MMSys*, Oslo, Norway, 2013.

[10] J. Schantl, C. Wagner, R. Kaiser, and M. Strohmaier. The utility of social and topical factors in anticipating repliers in twitter conversations. In *ACM Web Science (WebSci2013)*, 2013.

[11] O. Schreer, I. Feldmann, C. Weissig, P. Kauff, and R. Schäfer. Ultrahigh-resolution panoramic imaging for format-agnostic video production. *Proceedings of the IEEE*, 101(1):99–114, 2013.

[12] M. Steen and I. van de Poel. Making values explicit during the design process. *IEEE Technol. Soc. Mag.*, 31(4):63–72, 2012.

[13] I. T. Union. *One-way Transmission Time: Recommendation G.114 (05/03)*. ITU-T recommendations. ITU, 2003.

[14] M. F. Ursu, M. Falelakis, M. Groen, R. Kaiser, and M. Frantzis. Experimental Enquiry into Automatically Orchestrated Live Video Communication in Social Settings. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX'15, pages 63–72. ACM, 2015.

[15] M. F. Ursu, M. Groen, M. Falelakis, M. Frantzis, V. Zsombori, and R. Kaiser. Orchestration: Tv-like Mixing Grammars Applied to Video-communication for Social Groups. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM'13, pages 333–342. ACM, 2013.

[4] `http://icosole.eu/`