Virtual Director Technology for Social Video Communication and Live Event Broadcast Production

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

ABSTRACT

This thesis investigates several aspects of Virtual Director technology, i.e. software capable of intelligent real-time selection of live media streams. It addresses several research questions in this interdisciplinary field with respect to how a generic Virtual Director framework can be constructed, and how its behavior can be modeled and formalized to realize professional applications with many parallel users within real-time constraints. Prototypes have been built for the applications of group videoconferencing and live event broadcast. The engine executes cinematic principles aiming to enhance the user experience. In group videoconferencing, a Virtual Director aims to support communication goals by selecting from multiple available streams, i.e. automating cuts between shots according to the communication situation. In event broadcast, it enables personalization by framing, animating and cutting virtual camera views as cropping from a high-resolution panorama. While the technical approach and framework has been evaluated in lab experiments, further evaluation involving potential users and cinematic professionals is ongoing.

Categories and Subject Descriptors

C.2.4 [**Distributed Systems**]: Distributed applications; H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Virtual Director, live broadcast, video communication

1. VIRTUAL DIRECTOR TECHNOLOGY

The increase of services based on live video streams is very visible to users through a high rate of new multimedia applications with ever-improving audiovisual quality. New forms of content personalization can be enabled by intelligently cutting live AV sources. Since such personalization can not be executed by human operators on a large scale, intelligent *Virtual Director* software metaphorically aims to replace the work and knowledge of a live TV broadcast team. This concept is a key enabler for immersive experiences on top of that, and might enable future research advancements such as novel media formats and consumption paradigms.

This work explores the concept of Virtual Director systems and aims at research progress towards flexible re-usable software frameworks that can be adapted to a specific purpose by exchanging the production grammar, the Virtual Director's formalized behavior. This work reveals and addresses a set of research challenges regarding both the software framework and behavior definition aspects. Which technical approaches exist for implementing such software, what are their advantages and limitations? How could a generic Virtual Director framework be constructed that is capable of realizing professional large-scale applications, with a high number of parallel users, and tight real-time constraints? How can the execution of cinematographic principles in real-time media stream selection be automated? How can complex sets of Virtual Director behavior be modeled and formalized? How can it be designed such that it supports the users' aims best?

This paper discusses several aspects of this work on Virtual Director technology for the distinct domains of video communication and broadcast production. Before practical goals such as engineering a concrete Virtual Director behavior on an aesthetic level that meets user expectations can be taken on, underlying technical challenges need to be addressed. While this work builds to a certain extent on research using recorded multimedia content (for example [15]), it entirely focuses on applications using real-time live streams and their specific challenges, for example regarding delays in content transmission and user interaction. The requirements for a Virtual Director application in real-time scenarios are very challenging, and considerably different from remixing or abstracting recorded content. In this work, the audio-visual aesthetics and quality of automated cinematic principles are essential, unlike e.g. many surveillance applications based on different criteria and strictly professional users.

Even though the two domains of mediated communication and personalized event broadcast are significantly different, a common technical approach is sought. Each of these domains has specific requirements regarding the expected QoE delivered to the users, however, and each concrete production/setup requires a (domain-)specific set of Virtual Director behavior.

The focus so far was on understanding the concept's potential and limitations, and on developing a suitable tech-

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502213.

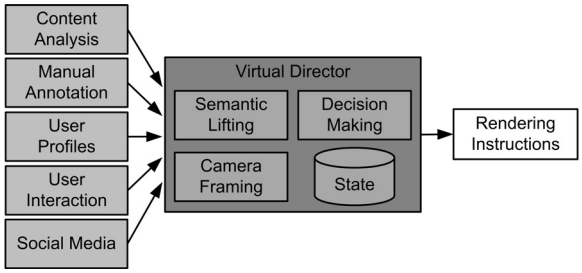

Figure 1: Generic workflow and sub-processes of a Virtual Director.

nical approach with the long-term research goal of creating a flexible Virtual Director software framework and behavior engineering methodology in mind. Software prototypes for both aforementioned application domains have been developed and applied. The prototypes execute dedicated Virtual Director behavior and are integrated with related system components in the realm of three larges research projects.

One of the thesis' core contribution is work on *semantic lifting*, i.e. on the semantic abstraction of low-level real-time information streams in order to inform the decision-making processes. In other words, this refers the understanding of the situation covered by the Virtual Director, i.e. the communication actions and patterns in one application domain, the actions going on in the scene in the other. Many approaches have been proposed for the detection of interesting events in video streams, e.g. a visual stream approach in [21]. An overview has been compiled recently in [7]. A Virtual Director's semantic lifting process however is typically operating at a very abstract level and has a slightly different aim, since it is triggering cinematic rules as a direct reaction.

Evaluation regarding both domains has been conducted via lab experiments and interviews, and initial results have been published as evaluation work is ongoing. While the Virtual Director concept is certainly bounded by practical limitations regarding the automation of human intuition, anticipation and creativity, several application domains are obvious candidates to exhibit added value to the users, most importantly through increased interaction possibilities and novel forms of personalization in the realm of live AV streams.

2. APPROACH AND ARCHITECTURE

Multiple technical approaches have been investigated in order to address the challenge of a flexible Virtual Director system that is both capable to reason with its information sources in real-time, and to take meaningful decisions based on them. One option is to represent the Virtual Director behavior in RDF/OWL, and to utilize Description Logics based reasoning and (Jena) rules. While the approach is well suited for static modeling of a production domain, experiments resulted in the conclusion that available reasoning frameworks are not fast and scalable enough for the requirements envisioned – see [9]. Ultimately, a rule-based approach was selected in combination with event processing technology [10], i.e. *Complex Event Processing* (CEP) [3] engines.

An abstract Virtual Director workflow is depicted in Figure 1. Such systems need to be informed about the content they cover in order to take meaningful decisions. This can be achieved through automatic content analysis (reported in [8]), manual annotation via dedicated user interfaces, or external sources such as the Internet.

A Virtual Director can be regarded as a software framework that executes a set of production grammar, its intended behavior. On a deeper level, the sub-processes deal with the distinct challenges of *understanding* the current situation, and of taking appropriate decisions based on that. While the concrete architecture depends on the intended personalization capabilities and user interaction requirements, three generic sub-processes of a Virtual Director are:

- **Semantic Lifting:** *Understanding* the scene or situation by semantically abstracting the low-level cue stream the Virtual Director is receiving through its sensors. Events and actions need to be detected that are on an abstraction level that allows to trigger Virtual Director behavior.
- **Camera Framing:** In order to achieve high QoE, cinematographic principles shall be applied. One key aspect is how to frame both static and moving shots. This is especially relevant if cropping from high-res cameras are utilized.
- **Decision Making:** Since the core added value of Virtual Director technology is grounded in the ability to personalize in real-time, efficient decision making is important. In contrast to a human production team, a Virtual Director can take a large amount of decision in parallel, enabling to take personalized decisions for every user or playout device.

The following Sections 3 and 4 discuss work in the two selected application domains in more detail.

3. GROUP VIDEOCONFERENCING

Enhancing social group communication through an *orchestrated* audiovisual link is a multifaceted and interdisciplinary research challenge. Within the European research projects TA2[1] and Vconect[2], Virtual Director prototypes have been built and evaluated [4, 18]. The so-called *Orchestration Engine* aims to support the communication goals of larger groups linked by an audio and video connection. As too many media streams are available (cp. [19]), this Virtual Director selects those most useful and enjoyable to the participants in order to support individual communication goals. Usefulness was measured by measuring the users' effectiveness in a task/game [6]. Experiments found that it does also influence the communication itself through its decisions. Support for validity of the concept has resulted from initial manual camera selection experiments [18].

The Virtual Director's role can be compared to human directors in live TV broadcast, calling the shots in a video conference in that case. Here, real-time constraints are even more crucial, since content delay affects bidirectional communication strongly. Social videoconferencing platforms for informal group communication vary in setup, using multiple microphones, loudspeakers, cameras and screens. Besides other social communication goals, one aim is to achieve a *telepresence* effect, i.e. to lose the feeling of communicating from distant locations.

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

Figure 2: Simultaneous virtual cameras as crops from high resolution panoramic video. The Production Scripting Engine frames shots based on its understanding of what is happening in the scene, as suggested by its semantic lifting sub-process.

Intelligent shot selection on that end aims to make sure the most relevant communication partners are visible large enough and at the right time. A high-quality full screen close-up shot for example allows to grasp also non-verbal communication cues well, which seems to be important especially for observing the communication partners reactions.

The domain is strongly related to business videoconferencing solutions such as Cisco's, but aims to support different communication processes and values (agenda-based, non-verbal communication less important). Videoconferencing has been looked into from many research angles. The use of multiple cameras to capture a person compared to current popular solutions like Skype and Google Hangouts has also been proposed by [20], but is available only in professional videoconferencing systems so far. While a lot of work has been reported with respect to eye gaze improvements and even 3D content capture (see [12]), this work is focusing on automatic shot selection.

One key advantage of being able to choose between multiple cameras that was investigated as part of this work is to select the most suitable camera view depending on the communication situation. While a frontal camera view might be most useful when speaking to someone directly, a dialogue between third parties is better shown from side angles. More detailed experiment results are reported in [18] and [10] and [6].

Regarding the semantic lifting sub-process, our current approach is to aim for understanding communication patterns, with turn taking events and patterns as the core concept to trigger camera selection decisions [5]. While both audio and video content analysis results are processed as real-time event streams, only the visual output is steered by the engine so far. To personalize dynamic content selection, static or dynamic information about the users may be exploited. From that end, we started analyzing communication behavior of users in social media, with the ultimate goal to predict conversation patterns such that they can be applied to camera selection in the video conferencing domain [16].

A future aim is to support dynamic communication topologies, intelligently mediated branching of side-conversations, which means that a sub-group from a communication session shall be able to create a separate communication space and come back to the main branch later. Another current focus is the integration of social videoconferencing with Social Network platforms, and the integration of the social group communication process with shared media consumption.

4. BROADCAST PRODUCTION

In a second application domain addressed by the FascinatE[3] project, a Virtual Director for live event broadcast [11] has been implemented, based on a scene-capture approach [13] using multiple microphones and a 180 degree panoramic camera, the OMNICAM [17]. See Figure 2 for an example frame with virtual cameras. The system automatically produces individual content streams for different playout devices and user preferences, i.e. mostly static or animated cropping from the high-resolution panoramic video. Viewers may watch different parts of the scene, based on their interests. This Virtual Director is informed by automatic content analysis (e.g. reported in [8]) and manual annotations. Manual input indicates the occurrence of high-level events not directly observable by content analysis, and manages virtual cameras.

For the prototype, called *Production Scripting Engine,* camera selection behavior was designed and implemented for the domains of soccer and dance performance, and different manual annotation strategies have been tried out to inform it. While high-level actions could be observed well by human operators in the soccer domain, the annotation of virtual camera candidates proved to be more useful in the dance domain, where few high-level concepts could be found that appeared to be intuitive enough to observe and annotate.

This work extends previous efforts by other researchers, many of which chose to work with sports content. Most work, however, is processing the media content in recorded form, not in real-time. Impressive results have been achieved within the APIDIS project (see e.g. [1]), though in different setups than this work. They utilized and registered multiple cameras in order to achieve stable player tracking for basketball matches, which is not possible to achieve when working with a limited set of camera angles, as conclusions and other issues occur. Daniyal and Cavallaro [2] also developed an algorithm for automated camera selection and used a Turing test to evaluate the quality of the output. Patrikakis et al. [14] aimed to create context-aware and personalized media in real-time streaming environments for broadcasting applications. Their approach is to allow users to create their own Virtual Director and to exploit the user profiles for camera selection decisions.

Evaluation confirmed that the quality of the visual decisions of the current prototype cannot compare with profes-

[3] `http://www.fascinate-project.eu/`

sional production teams and is lacking in perceived creativity, storytelling skills and intuition. Still, reasonable application areas exist. A Virtual Director provides the advantage of quicker reactions to low-level cues, and consistency in decision-making that is hypothesized to result in a more reliable experience. The main advantage of Virtual Director technology for live event broadcast, however, are the manifold possibilities to personalize camera selection to the benefit of the user. The concrete options depend on the production domain, and need to be made available to the users through an interface.

In initial user evaluations, we found that user expectations depend on the production domain. For some they are very high, as they are based on experience consuming professionally scripted and edited TV programmes. Still, both production professionals and unskilled users participating in evaluation interviews were able to ignore superficial deficiencies to understand and acknowledge the personalization opportunities enabled by such technology.

One option to circumvent some of the inherent quality limitations of the concept would be to go for director decision assisting instead of automatic decision making as a more short-term goal. While this approach makes sense in certain broadcasting applications, it could not be applied to the videoconferencing domain, where human professionals cannot be involved for economic reasons.

5. STATUS AND OUTLOOK

An efficient technical approach has been proposed that allows to build Virtual Director systems and to define their intended behavior. Virtual Director technology, however, remains a challenging research area.

This works' core aim is to advance towards the long-term goal of a generic software framework, a proper formalism and authoring tools for engineering Virtual Director behavior. Its goal is not to engineer a system with a perfect set of behavior for a certain application. Since the prototypes' practical limitations can not be disregarded in user evaluations, the evaluation of the concept is conducted with respect to several individual aspects as discussed above, rather than as direct comparison to e.g. a TV broadcast edited by a professional director. Quantitative results have been obtained evaluating both system performance and user goals.

While Virtual Director prototypes have been implemented successfully based on this approach, a key challenge of the approach is how to deal with competing and contradicting principles that need to be resolved dynamically for decision making [10]. Further future work is to put focus on steering audio playout, since so far mainly the visual domain had been addressed.

6. ACKNOWLEDGMENTS

The research leading to this paper has been supported by the European Commission under the contract FP7-287760, "Vconect - Video Communications for Networked Communities", and FP7-248138, "FascinatE – Format-Agnostic SCript-based INterAcTive Experience".

7. REFERENCES

[1] F. Chen, D. Delannay, and C. De Vleeschouwer. An autonomous framework to produce and distribute personalized team-sport video summaries: A basketball case study. *IEEE Transactions on Multimedia*, 13(6):1381–1394, 2011.

[2] F. Daniyal and A. Cavallaro. Multi-camera scheduling for video production. In *In Proceedings of the 9th European Conference on Visual Media Production*, pages 11–20, 2011.

[3] O. Etzion and P. Niblett. *Event Processing in Action*. Number ISBN: 9781935182214. Manning, 2010.

[4] M. Falelakis, M. Groen, M. Frantzis, R. Kaiser, and M. Ursu. Automatic orchestration of video streams to enhance group communication. In *SAM'12 Workshop at ACM MM*, pages 25–30. ACM, 2012.

[5] M. Falelakis, R. Kaiser, W. Weiss, and M. Ursu. Reasoning for video-mediated group communication. In *ICME*, 2011.

[6] 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.

[7] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, pages 1–29, 2012.

[8] R. Kaiser, M. Thaler, A. Kriechbaum, H. Fassold, W. Bailer, and J. Rosner. Real-time person tracking in high-resolution panoramic video for automated broadcast production. In *CVMP*, pages 21–29, London, UK, 2011.

[9] R. Kaiser, C. Wagner, M. Hoeffernig, and H. Mayer. The interaction ontology model: supporting the virtual director orchestrating real-time group interaction. In *MMM'11*, pages 263–273, Berlin, Heidelberg, 2011. Springer-Verlag.

[10] R. Kaiser, W. Weiss, M. Falelakis, S. Michalakopoulos, and M. F. Ursu. A rule-based virtual director enhancing group communication. *ICME'12 Workshops*, pages 187–192, 2012.

[11] R. Kaiser, W. Weiss, and G. Kienast. The Fascinate Production Scripting Engine. In *MMM'12*, pages 682–692, Berlin, Heidelberg, 2012. Springer-Verlag.

[12] A. Maimone, J. Bidwell, K. Peng, and H. Fuchs. Enhanced personal autostereoscopic telepresence system using commodity depth cameras. *Computers & Graphics*, 36(7):791–807, 2012.

[13] 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'13*, Oslo, Norway, 2013.

[14] C. Patrikakis, N. Papaoulakis, P. Papageorgiou, A. Pnevmatikakis, P. Chippendale, M. Nunes, R. Cruz, S. Poslad, and Z. Wang. Personalized coverage of large athletic events. *IEEE MultiMedia*, 18(4):18–29, 2011.

[15] L. A. Rowe and V. Casalaina. Capturing conference presentations. *IEEE MultiMedia*, 13(4):76–84, 2006.

[16] J. Schantl, C. Wagner, R. Kaiser, and M. Strohmaier. The utility of social and topical factors in anticipating repliers in twitter conversations. In *WebSci*, 2013.

[17] 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.

[18] M. 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 *ACM MM'13*, New York, NY, USA, 2013. ACM.

[19] M. Ursu, P. Torres, V. Zsombori, M. Franztis, and R. Kaiser. Socialising through orchestrated video communication. In *ACM MM'11*, pages 981–984, New York, NY, USA, 2011. ACM.

[20] P. N. Y. Wang and M. Kankanhalli. Multi-camera skype: Enhancing the quality of experience of video conferencing. In *PCM'11*, pages 193–202, 2011.

[21] B. Yu. *MyView: Customizable Automatic Visual Space Management for Multi-Stream Environment*. PhD thesis, University of Illinois at Urbana-Champaign, 2006.