Vconect - Orchestration for Group Videoconferencing

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

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

Manolis Falelakis
Department of Computing
Goldsmiths, University of London
London, UK
m.falelakis@gold.ac.uk

ABSTRACT

Current videoconferencing systems show a lack of support in adapting visual remote camera presentation to the users' needs. Some manage to put focus on the current speaker. In this demonstration we show an automatic decision making component in the realm of social video communication that aims to go beyond that. Our approach takes into account several aspects such as the current conversational situation, conversational metrics of the past and device capabilities to make decisions on the visual representation of available video streams. This allows to optimally support users in communication within various communication contexts.

Author Keywords

Videoconferencing, communication orchestration, automatic decision making

INTRODUCTION

Audio-visual communication pervades slowly but continuously our daily life, mainly driven by the availability of broadband services and mobile devices. One important aspect in our life is social communication with our friends. This type of communication is featured by certain characteristics for example people could join and leave continuously, the dynamic of the conversation might change over time and the network capabilities might change. Current video communication systems for social communication have limited intelligence to adapt to specific communication situations. We argue that taking into account conversational metrics and other parameters such as device capabilities and to adapt the visual representation of the videoconferencing client accordingly helps the user to get immersed by the communication experience.

The Vconect[1] project investigates novel ways of supporting mediated audio-visual communication for ad-hoc groups. One 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 (no self-view assumed). The question is how to optimally deal with them. An intuitive, but not always scalable option, is to show each user all video streams side by side on one screen (referred to as tiled layout). We set out to investigate more sophisticated solutions with the aim of achieving better communication support through intelligent camera selection. When taking into account further parameters such as conversational metrics [1] which represent the dynamics of a conversation, and the capabilities and size of the client it is possible to select a suitable visual layout and the corresponding video streams. A component which automatically executes a mixing process of different video streams is known as a Virtual Director [2]. In the realm of communication this is mostly referred to as Orchestration [4].

Subsequently we discuss the architecture and influencing parameters of the decision making component which intelligently switches between visual representations and executes the video mixing process for each user. Then the demo is described in detail and highlights what users can expect.

ORCHESTRATION

Communication Orchestration is the decision making which controls the mixing process of all available video streams. It can be compared to compiling a live TV transmission but in the case of video conferencing it has to address communication rather than narrative needs. Orchestration is a reasoning process which operates in real-time for each location participating in the communication. It builds upon the audio and video processing infrastructure and executes camera control and audio-visual composition (cf. [4]).

The Orchestration Engine in the Vconect project automatically produces camera selections and visual layout changes by reasoning on audio-visual cue streams from its participants. The system is implemented as a three step process:

Cue extraction: Audio-visual streams are processed by analysis modules in the system underneath and low level cues are extracted in real-time. An example for a low level cue is "voice activity".

Fusion and interpretation: Low-level cues from all locations are aggregated in the Semantic Lifting module of the Orchestration Engine. In this stage, higher-level semantic events that concern the communication as a whole, such as a "turn-shift", are generated while properties of the state of the interaction at each point, such as "active speaker"

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

2nd International Workshop on Interactive Content Consumption at *TVX14*, June 25, 2014, Newcastle, UK.

or "crosstalk", are evaluated continuously. The module aims to achieve a computational interpretation of the current communication situation on a semantic level that can directly be evaluated by the decision making components. The Semantic Lifting module also calculates conversation metrics such as "turn shifts per active participants" or the "active participation duration per each participant" based on a sliding temporal window. These conversation metrics allow to identify monologue situations or to identify the "heatedness" of a discussion.

Decision making: The application of mixing rules that result in the shot selection and the selection of the visual layout is made by the Director modules. For each screen one separate instance reasons based on high-level events and conversation metrics received from Semantic Lifting and other modules. This process allows to select the optimal visual layout in combination with the necessary video streams for each user and gives best support in communication by respecting the given limitations.

The two latter components are part of the Orchestration Engine which is a central, server-side software component. The logic is implemented declaratively using forward-chaining rules of the JBoss Drools[2] reasoning engine. Detailed examples of rules already incorporated into the system together with challenges about their implementation were reported in [3].

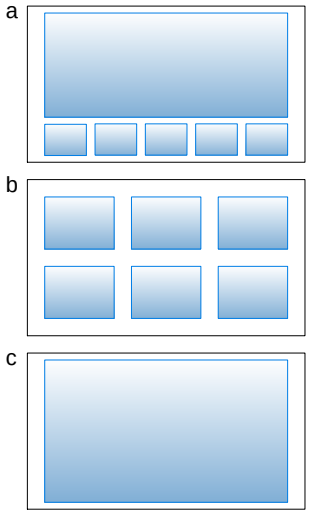

Figure 1. (a) Layout with focus on one person and with smaller tiles for all other participants. (b) Tiled view layout. (c) Full screen layout.

THE DEMO

This demo shows the abilities of a video conferencing system for social communication that takes into account various system parameters as well as conversation metrics to select the optimal visual layout for each participant. Figure 1a illustrates a layout with focus on one person and with smaller tiles for all other participants. This is suitable for most situations when the screen size is big enough and the number of participants does not exceed more then 10 people and the conversation is in a low or normal pace. If the conversation gets more heated (animated), meaning there is a high number of turn shifts between the participants within the analysis time window, it would be beneficial for the user to see all involved participants on the screen. This visual layout is illustrated in figure 1b but this can only be applied if certain other parameters allow a switch e.g. there needs to be enough space on the users screen. A single full screen layout which has only one video stream (see figure 1c) will be chosen when there is a monologue detected by a participant. Another reason for a single full screen layout would be if the the screen size is very small.

Interested persons have the possibility at the live demo to use one videoconferencing node on site to join a video conferencing session together with four remote participants. The remote participants are located in different offices in different countries. Video streams are transmitted in HD quality from all participants. Users can experience a videoconferencing system which automatically selects a suitable visual layout and the right video streams to optimally support users in communication. It is also possible to manually select the visual layout so that the automatic decision making system can be easily compared.

[2]https://www.jboss.org/drools/

ACKNOWLEDGMENTS

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. We thank all partners who contributed to the Vconect system for their support to make this demo possible.

REFERENCES

1. Hammer, F., Reichl, P., and Raake, A. The well-tempered conversation: interactivity, delay and perceptual voip quality. In *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 1 (May 2005), 244–249 Vol. 1.

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

3. Kaiser, R., Weiss, W., Falelakis, M., Michalakopoulos, S., and Ursu, M. A rule-based virtual director enhancing group communication. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on* (July 2012), 187–192.

4. Ursu, M. F., Groen, M., Falelakis, M., Frantzis, M., Zsombori, V., and Kaiser, R. Orchestration: Tv-like mixing grammars applied to video-communication for social groups. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, ACM (New York, NY, USA, 2013), 333–342.