Automatic Orchestration of Video Streams
to Enhance Group Communication

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

Martin Groen
School of Experimental
Psychology
University of Bristol
Bristol, UK
martin.groen@bristol.ac.uk

Michael Frantzis
Department of Computing
Goldsmiths,
University of London
London, UK
m.frantzis@gold.ac.uk

Rene Kaiser
Institute of Information and
Communication Technologies
Joanneum Research
Graz, Austria
Rene.Kaiser@joanneum.at

Marian F Ursu
Department of Computing
Goldsmiths,
University of London
London, UK
m.ursu@gold.ac.uk

ABSTRACT
Unlike legacy video-conferencing, which connects two nodes each equipped with a camera, recent systems facilitating for video-mediated group communication deal simultaneously with a large number of video streams. This highlights the need for orchestration, i.e., the intelligent selection of the most adequate camera views to be displayed on each screen. In this paper we present the initial results of a study that evaluates the effects of orchestration on communication within a specific context; that of two remote groups playing a collaborative board game. The results of the experiment indicate that automatic orchestration can provide improvements similar to the ones achieved when live video mixing is performed by human editors.

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

General Terms
Design, Experimentation, Human Factors

Keywords
Orchestration, Rule-based reasoning, Task performance, Video-mediated communication

1. INTRODUCTION
Video-conferencing has penetrated the informal space of social communications, and has developed into a tool that many people use on a regular basis (Skype reported 663 million users in 2010) and increasingly so by geographically separated families [1]. However, the intelligence of those systems to adapt to specific communication situations is very limited and they display significant limitations related to the setups in which they can be used and the communication topologies they can support. In particular, they are very poor at supporting group communication and allowing people to move in their respective physical spaces. Recent commercial products start to provide simple solutions to some of these limitations by providing a selection of multiple simultaneous video streams. Multipoint Skype and Google+ Hangouts allow multiple spaces to be connected [13], each using one camera, whereas Polycom Eagle Eye Director [11] is able to provide close-ups of active people in the space by using an extra pan-tilt-zoom camera.

In both cases, the number of available streams can easily grow very large, as the amount of spaces and/or cameras grows, and displaying them all as a mosaic is not an option. This raises the need for communication orchestration, i.e., the automatic selection and composition of the most appropriate camera views to display on each individual screen [14].

Face-to-face communication, which provides people with unhindered access to the full set of verbal and non-verbal communication channels, is more effective than mediated communication as media introduce restrictions on the ability of people to utilise fully all communication channels [2,5]. A significant aim of successful multi-camera orchestration is to make the medium transparent allowing those communicating to see and hear each other in manners similar to collocated experiences.

Recent studies [6] have demonstrated that orchestration can improve video-mediated communication over "traditional" static shot approaches. Video mixing in these experiments was carried out manually, i.e., by employing human editors. However, orchestration is a process that inherently has to be carried out automatically.
For an intelligent software, i.e. an Orchestrator to undertake such a task, it needs to automatically identify which of the numerous actions or events happening simultaneously is the most relevant for each individual location, by adhering to an appropriate set of mixing rules. This process relies essentially on efficient audiovisual scene analysis and cue extraction (see [10] for a survey of related methods) and includes real-time processing of events [3] and spatio-temporal reasoning. The architecture of such systems has been recently presented [4], however without experimental evidence.

In this paper we present the initial results of a study that examines the effects of automatic orchestration by carrying out an experiment involving two groups of people that play a collaborative board game while located in two remote locations, each covered by three cameras. In order to assess the effect of orchestration we compare the participants’ task performance under four conditions: (i) automatic orchestration, (ii) manual orchestration, (iii) no orchestration (static wide shot) and (iv) random mixing. The last condition serves as a baseline, helping to measure the effect of informed decision making compared to plain alteration of the visual scenes.

This experiment was hypothesized to show that the Orchestrator can achieve (at least) similar improvements to the participants’ ability to achieve their communication and collaboration goals as the ones resulting by human editing.

In the next section we briefly present the functionality of the used system, while section 3 is devoted to the description of the experimental design and setup, together with the adopted orchestration rules. Section 4 presents the results, which are discussed in section 5. Finally, section 6 concludes the paper with plans for the continuation of this work.

2. SYSTEM DESIGN

Our system is an instantiation of our previous work [4], which automatically produces camera selections by reasoning on audiovisual streams from the participating rooms. Inference is based on a set of declarative rules, the execution of which is essentially a 3-step process, as depicted in Figure 1:

- **Cue extraction**: Audiovisual streams are processed and low level cues are extracted. This is done on the fly by a series of **Analysis** modules, one per room/ location. Video is provided by a wide camera capturing activity in the room and sound by an array of directional microphones, while the process of cue-extraction is described in [9]. Examples of such cues are "voice activity" and "face position".

- **Fusion and interpretation**: Low-level cues from all locations are aggregated in the **Cue Lifter** module. 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" or "cross-talk", 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.

- **Decision making**: The application of mixing rules that result in the shot selection is made by the **Director** modules. Although multiple modules run in parallel (one per screen), they do not operate in isolation; their decisions are based on high-level events and state information received from the Cue Lifter and originating from all locations.

The two latter components are part of the Orchestrator, which is implemented using forward-chaining reasoning of declarative rules using the JBoss Drools reasoning engine [7]. Detailed examples of rules already incorporated into the system together with challenges and workarounds in their implementation were reported in [8].

3. EXPERIMENT

3.1 Participants and design

We carried out fourteen 2-hour sessions, each involving 4 people, thus employing a total of 56 participants. The data of 12 of them had to be dropped, due to technical problems or no-show, leaving 44 participants whose age varied between 18 and 51.

Four conditions were created:

1. **Automatic Orchestration**: the selection of shots for each of the two locations was done automatically by the Orchestrator.
2. **Manual Orchestration**: two human editors, one per location, selected the shots from the multiple camera streams.
3. **No Orchestration**: a single camera displayed video imagery on the TV sets consisting only of a wide shot of the other room.
4. Randomly edited: this condition was included to check that any effects found are not confounded by the effects of changes in imagery on the TVs alone. Random mixing was achieved by alternating the same shots that were available for the orchestrated condition, but in a random manner and in random intervals.

There were two teams of two participants in each session who experienced the four different conditions in succession, giving a three factorial within-participants design. The participants were unaware of the different conditions, each of which had a duration of 15 minutes and was succeeded by filling out a questionnaire. The order of the conditions was counterbalanced so as to reduce carryover effects, such as variations in game performance because of practice, assimilation, fatigue and/or boredom.

Dependent variables were the number of correctly guessed words in an Articulate game and the responses to a subjective experience questionnaire.

3.2 Setup

An audio/video connection was implemented between two distant rooms that were each equipped with a HD 50 inch TV and three high-quality HD cameras positioned centrally at the base of the TV.

![Figure 2: Room layout and camera positioning.](image)

The room layout is roughly depicted in Figure 2. The central camera (WIDE) provided a wide-shot covering the room, while the other two cameras (CU-LEFT and CU-RIGHT) provided frontal close-ups on each of the two persons in the room. There was also a coffee table in front of the participants on which the card games and hour glasses were placed, together with refreshments.

The scenario involved 2 teams playing a game of Articulate across the video link. Articulate is a talking description board game, where players describe words chosen from a pile of cards to their team mate, as quickly as possible, within 30-second rounds. Four participants played in teams of two and team members were separated across the video link. Players were allowed to use verbal descriptions as well as gestures.

3.3 Orchestration rules

In order to provide meaningful comparison between conditions, we had to define a closed common set of rules to guide the decision of both orchestrated conditions. This has been essential as we wanted to prevent the editors from relying on the implicit knowledge provided by their expertise. Even though our system allows for the implementation of richer rules with complex interplay, for this experiment we decided to devise a restricted principled procedure which would ideally be (i) relatively small and simple, so that it could be followed by humans with consistency and at the same time (ii) rich enough to cover the ongoing reaction both pragmatically and aesthetically.

To achieve this we based the editing decisions on the notion of turn-taking in conversation [12]. The mixing rules were:

1. If person P starts a turn (i.e. a turn-shift to P occurs), show a Close Up (CU) on P.
2. If there is a pattern of short turn-taking (more than 3 turn-shifts within 5 seconds) in the same room then show WIDE.
3. If there is no turn-shift for 5 seconds, then show WIDE.
4. No shot change is allowed if no more than 2 seconds have passed from previous change.

Whilst the execution of these rules is realized directly by a human editor, when carried out by the Orchestrator it involves three stages, corresponding to the ones described in section 2. For instance, execution of rule number 1 begins with the detection by Analysis of a set of scattered voice activities and their association with a specific person. These voice activities are sent to the Cue Lifter, which decides that they constitute a turn-shift (excluding potential interjections). The Director is then informed and eventually executes the rule and sends the command for a shot change. Lists of cue lifting and directing rules used for this experiment are provided in tables 1 and 2 respectively, in a pseudo-language resembling first order temporal logic.

We must comment here that the actual rule execution is a much more complicated process, as a result of the abundance of events occurring during an interaction that simultaneously trigger competing and conflicting rules. A clear example of this situation is rule number 4 which stops other rules from being executed if no more than 2 seconds have passed after the last camera change. Careful implementation of the reasoning mechanism and rule interplay is required to achieve conflict resolution [8] but this is beyond the scope of this paper.

In order to assist the human editors and to ensure highest consistency possible, we enforced rule number 4 by incorporating a corresponding timer in the live-mixing software.

4. RESULTS

The acquired quantitative results are depicted in Figure 3. Average game points per session are shown by experimental condition, together with the corresponding standard error represented with error bars.

The main effect of orchestration shows that it affected the game play of the participants, $F(3, 30) = 4.34, p = .01, \eta^2_p = .30$. Mean number of game points won in the automatic condition was 25.91 (SD = 8.07), in the manual condition 24.00 (SD = 9.74), in the No Orchestration condition mean game points won was 23.09 (SD = 11.42) and in the random video mixing condition mean game points won was 17.27 (SD = 9.03). Planned pairwise comparisons, Bonferroni corrected, showed that only the Automatic orchestration condition contributed to the observed difference in effect of orchestration, $p = .05$.  

27
TABLE 1: List of cue lifting rules

1. \( \text{voiceActivity}(P) \land \lnot \text{turn}(P) \land \lnot \text{crossTalking}(P) \rightarrow \text{turnShift}(P) \)

2. \( \text{numberOf}([\text{Turnshift}(P_1) \lor \text{Turnshift}(P_2)])d(t, t - X) > Y \rightarrow \text{patternOfShortTurnTaking}(P_1, P_2) \)

TABLE 2: List of directing rules

1. \( \text{turnShift}(P) \land \lnot \text{isInLocation}(L, P) \rightarrow L: \text{cut2CUFront}(P) \)

2. \( \text{patternOfShortTurnTaking}(P_1, P_2) \land \lnot \text{isInLocation}(P_1, P_2) \rightarrow \text{cut2Wide}(P_1) \)

3. \( \text{timeSinceTurn}(P_1, P_2) > \text{threshold} \land \text{activePerson}(P) \land \lnot \text{isInLocation}(P_1, P_2) \rightarrow \text{cut2Wide}(P) \)

4. \( \text{durationOfCurrentShot} < 2\text{sec} \rightarrow \text{forceMaintainShot} \)

Figure 3: Average game points won by experimental condition.

5. DISCUSSION

The results suggest that our hypothesis was corroborated as participants demonstrated similar improvements when collaborating in both the orchestrated conditions. These improvements might have been facilitated by the availability of more useful information from their social environment. Furthermore, only the automatically orchestrated condition did achieve significant improvements with respect to the baseline, constituting an important finding and showing the potential of the approach.

Another observation is that the orchestrated conditions did not demonstrate significant improvements when compared to the static condition, as reported in our earlier work [6]. This can be attributed to the fact that, whereas that study involved a selection of 5 different shots per location, this one utilized only 3, reducing the potential improvement due to proper selection.

We also need to point out that, however much restrictive the imposed set of rules, human editors naturally benefit from their implicit knowledge and understanding of the conversational context and a rich set of verbal and non-verbal cues. The Orchestrator is handicapped in this sense, having to work with the inherent limitations of audiovisual analysis, in terms of number of features and detection accuracy, and the closed set of rules.

On the other hand, the Orchestrator provides the advantage of much quicker reactions to 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. That is, due to factors such as fatigue, boredom, difficulty in hearing and seeing the events in the room, and inherent differentiation in human mixing responses to salient events, human mixing presents a less reliable orchestration experience as the decisions, almost by necessity, will be different across and within communication sessions. Automatic orchestration does not have this limitation (that is, a machine ceteris paribus always responds the same). The limitations that have been signaled are problems of a technical nature that can be solved in time. This is the important contribution of this work.

6. FUTURE WORK

Analysis of results is still ongoing. Our first priority will be to analyze the responses to the subjective evaluation questionnaire, which will hopefully provide more evidence and intuition.

We also intend to process the recordings in an attempt to assess the degree up to which the human editors followed the orchestration rules, or relied on their expertise. The same will be attempted for the automatic module where deviations will occur due to inherent imprecisions in the detection of cues and the definition of the rules.

Further quantitative statistics such as average response times in each condition, together with average number of turns, average length of turns and deictic gestures are expected to shed more light in the explanation of the observed effects and hints for further developments.

Our mid-term plans include the evaluation of the orchestrated experience in a three-room layout, including more cameras per location and more complicated sets of orchestration rules. We expect that these experiments will further reveal the potential of orchestration.

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.

7. REFERENCES


