Socialising through Orchestrated Video Communication

Marian Ursu  Pedro Torres
Vilmos Zsombori  Michael Frantzis

Department of Computing
Goldsmiths, University of London
{m.ursu,p.torres,v.zsombori,
m.frantzis}@gold.ac.uk

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

1. INTRODUCTION

We are working towards the development of a video-based virtual communication and interaction space in which dispersed groups of people are able to naturally talk to each other, and to create and share moments of fun, by engaging in entertaining social activities. In simple terms, this interaction mode can be described as a multi-location, multi-camera, hands-free video conference, with integrated support for engagement in social entertaining activities between groups of people who already know each other, such as family and friends. This work has been motivated in [8].

The aim is for the communication technology to create as natural and immersive a communication medium as possible between groups. Participants are free to move within their spaces and behave in ways close to a collocated unmediated interaction. Multiple cameras offer the benefit of different perspectives on each location and the system is responsible for providing the best support for the communication channels at each time during the interaction, seamlessly integrating the entertaining activities that are taking place. This process is called automatic orchestration. Orchestration is about developing new screen grammars for video-mediated communication and play and expressing it in a computational form amenable to automatic application during the interaction. The main requirements we chose include: (i) pragmatics — all that is important to be seen in each location is indeed shown; (ii) autonomy — absence of direct instructions to orchestration; (iii) transparency — end users’ perception of the communication medium should be minimal; (iv) aesthetics — screen storytelling grammar employed to better convey what is pragmatically required.

Work related to the overall paradigm has been the main topic of a EuroITV workshop[5]. Related research in interactive TV narratives investigates automatic creation of TV productions which adapt to viewers requirements at the time of delivery both with pre-recorded content [7] and live content [4]. A major distinction to our approach regards the roles played by the end users — actors vs spectators. Intelligent virtual camera planning [2] deals with the automatic planning of camera placement and movement, sometimes including also shot composition, in order to achieve the best narrative effect. Most of this work, although closely related, is carried out in virtual worlds, where there are fewer constraints regarding the cameras and far more information regarding the world itself. Extensions to the real physical space...
world exist [6], but they are quite limited in scope. Automatic orchestration of camera panning and zooming has been explored for video conferencing among children [9].

Research in social TV attempts to enhance the social experiences surrounding the watching of TV programmes, for example, by providing communication support for those separated in space and possibly time alongside the watching of the same TV content. Representative samples of recent research, including various approaches to this concept and surveys of the state of the art can be found in [1, 3]. The aim of establishing communication channels between people situated in different locations, with a view to sharing and enhancing the entertainment experiences is perfectly aligned with ours.

2. FORMALISATION

2.1 Architecture

The architecture of the Orchestration Engine (OE) is presented in Figure 1. The design approach is centralised — one engine for the whole system — and declarative — the knowledge bases are separate from the reasoning procedures. The OE receives primitive events from the ANALYSIS modules from all the locations, which analyse the local audio visual streams and extract low-level contextual information (e.g. voice activity from a particular person). The OE also receives input regarding the game state from the GAME ENGINE (e.g. whose turn it is or if there has been a winning move).

2.2 Interaction Ontology

The ANALYSIS modules identify communication features, in real-time, from the audio and visual streams of all locations. They are sent to the OE as timestamped events, e.g. voice(start, personID, timestamp). The GAME ENGINE also sends events that happen in the game space, e.g. winning_move(personID). We referred to both of these types of events as primitive.

The OE maintains a state of the overall social interaction and communication context, represented as a set of higher-level events, e.g. dialogue(personID, personID). The primitive cues are extracted from each location independently of the others, but the OE integrates them in the overall state. The OE recomputes the state each time it receives a new primitive event, on the basis of the definitions stated in an “interaction ontology”, a knowledge base currently expressed in OWL2. So far we have concentrated on classifying sequences of lower-level voice activity events into more meaningful conversational events, namely monologue, dialogue, group discussion and silence. To define them, we took only the start time of each speech into account, and neglect the duration, overlaps and gaps of each event. Whilst still quite simplified, this approach provided overall good results. Nevertheless, we encountered problems with regards to the efficiency of reasoning and a newer implementation of this cue lifting module which dropped the ontology approach in favour of Complex Event Processing is reported in [5].

2.3 Communication Channels and Rules

Channel inference rules map the high-level cues into channels of communication. These channels are captured by a dynamic three-dimensional matrix C with entries, Cijk, between 0 and 1. C is recomputed after each change in the state of the system. The value of Cijk is the weight of the channel from person i to person j as perceived by person k: 0 means no channel, whereas 1 means maximum importance. The need to introduce the third index, k, derives from the fact that a conversation between two people may be of different relevance to different observers. For example, if the channel inference module receives a dialogue event between person i and person j a rule may raise the weight of all channels Cijk for every k. The rise in weight could be greater if k = i or k = j and lower otherwise to express the fact that, when person i is talking to person j, the conversation is more important to person i and person j than to some other person. This dialogue rule could be stated as:

\[
\begin{align*}
    i = j & \Rightarrow C_{ijk} = 0 \\
    i \neq j & \Rightarrow \begin{cases} 
        k = i \vee k = j & \Rightarrow C_{ijk} = 1, C_{jik} = 1 \\
        k \neq i \land k \neq j & \Rightarrow C_{ijk} = 0.2, C_{jik} = 0.2
    \end{cases}
\end{align*}
\]

Figure 1: Architecture of the OE
Similar rules are used for silence, monologues, group discussion and unknown states. More generally, the inference rules may control the weights of channels of communication based on any kind of event raised via the IO rules. The contributions coming from different audiovisual states are all added into a single C matrix and subsequently normalised for each observer.

The support given by each camera to each channel of communication is captured by a four-dimensional matrix $S$ with entries, $S_{ijkc}$, between 0 and 1. The value of $S_{ijkc}$ is the support given by camera $c$ to the channel from $i$ to $j$ as perceived by $k$. The fitness of camera $c$ with respect to the location $L$ is defined as

$$\phi_L(c) = \sum_{ij} \sum_{k \in L} C_{ijk} \cdot S_{ijkc}. \quad (1)$$

The way camera fitness is defined implies that the choice of a camera over another is made considering superposition of partial fitmesses concerning different triples of people ($i, j, k$). Essentially, the fittest camera for a location is the one that provides the most support over all channels involving the observers in that location.

The matrix $S$ is determined using support rules. A support rule may state, for example, that, if person $i$ and person $k$ are in the same location — which implies they can see each other without the support of the system — and there is a channel of communication from $i$ to $j$, the support for observer $k$ of a camera showing a close up of person $j$ is 1. In other words, support will be 1 if $k$ can see both people involved in the dialogue. Support rules, in general, may consider the proportion a person occupies on the screen — seeing a wide shot with three people when we just want to see one is worse than seeing a close up of that person — and the angle at which they look at the camera — to cover a dialogue, a camera that shows a front shot is more useful than one that shows the person at an acute angle. For brevity, details of these rules are left out. According to our formulation, the support matrix $S$ may vary with time (e.g. if people move, different cameras will give different support) and the OE will make its decisions accordingly. This feature will be useful when there are cues providing information on real-time position of people and cameras.

Style rules aim both (i) to avoid too frequent or infrequent camera changes and (ii) to constrain transitions between types of shots. A style rule may state that a camera change is not allowed more frequently than every 2 seconds or that, after 60 seconds of showing the same feed, a different camera should be picked even if it is not the best choice. As for transitions, a style rule may state that a close-up of a person cannot be followed by a close-up of another person in the same room without going through a wide-shot of both people, to avoid disorientation. In summary, the OE computes the fitness of each camera with respect to every location according to $(1)$ and then chooses the most fit camera at each location that satisfies all the style constraints.

3. EVALUATION

The concept of orchestration was evaluated through three sets of experiments — Figure 2. In the first set of experiments, we worked with 8 participants who were organised in different combinations of groups of 2 and 3 and swapped between the two rooms. We fixed the following characteristics:

- (i) two locations; 2/3 people per location; (2) non-mediated audio link — the two locations were set adjacent to each other, but with no direct visual connection; (3) one TV set in each location; one room received a manually orchestrated video stream edited from 5 cameras from the other location; the other location received a static stream.

Figure 3 illustrates three shots available to the orchestration process to cover a group of friends playing Pictionary with other friends in another location. Cam$_A$ is a wide shot, used to cover both the sofa and the armchair, Cam$_B$ is used for showing the game board, but also for allowing the person at the board to communicate with the other location, and Cam$_C$ is used to show the person in front of the armchair when he becomes active. Besides having different roles in supporting communication, these shots also serve different aesthetic functions.

![Figure 2: Experimental Set-ups](image1)

![Figure 3: Example of functional shots](image2)
the participants. However, overall, the results strengthen our conviction that screen grammar has the potential to improve video-mediated communication between groups.

In a second set of evaluation experiments for automatically orchestrated communication we kept the dependency on ANALYSIS, but removed the requirement for real-time computation for all the software components, and also the reliance on the VCE. Instead, we used a non-linear editing suite which we called off-line automatic orchestration. For this, we carried out a one day trial involving 5 participants, work colleagues and friends, organised in different groups of 2. The set-up consisted of two separate rooms, communicating via mediated audio and video. The games played were Ludo and Tabu. For both rooms, the audio and video streams were static during the interaction, but one of the locations had three fixed cameras continuously recording, with the front wide camera being of sufficiently high quality to be used for virtual cameras. The real-time experience was not orchestrated, but automatically orchestrated streams were subsequently constructed and evaluated. ANALYSIS was carried out off-line on the recorded footage as a stand alone process. It produced a description of features synchronised with the audio-visual content. Orchestration was then run, taking these primitive cues as input. It produced a description of all the editing decisions and exported them as FinalCut-Pro XML files. The edited streams were then watched and evaluated by those participating in the experience as well as by others. The overall impression was again encouraging: the orchestrated stream was considered as more interesting and versatile than the static wide when the orchestration logic was complete — i.e. when it included interpretation of primitive cues into more meaningful information, the refinement of channels of communication, the use of camera support rules and, finally, stylistic rules.

The third and final set of experiments was carried out for real-time automatic orchestrated interaction. The set up consisted of two separate rooms, each having 1 central and 2 lateral cameras, the central one also supporting virtual cameras. The game played was an electronic version of the board game Space Alert. Apart from informal evaluation sessions carried out by the members of the team that developed the system, there have been two 90 minutes evaluation sessions, carried out by a group of 5 people. The initial ambition was to have all the processes running automatically. Unfortunately, the extraction of primitive cues in real-time was insufficiently reliable. This led to more or less random editing decisions and its effect was clearly disturbing. For example, at times, instead of showing the person talking, the system showed a close-up of the other person in the room. Therefore, to decouple the orchestration process from the unreliability of the cues, primitive cues were injected manually by a human operator observing the activity in the room. Orchestration using such primitive cues was deemed as interesting and with clear exploitation potential in video mediated communication.

4. CONCLUSIONS AND FUTURE WORK

We have described a framework for orchestrating video communication for social interaction between groups of people via a multi-location, multi-camera, hands-free system. To the best of our knowledge, this is the first attempt to develop a screen grammar for social interaction and the necessary formal models and associated tools to express and reason with it.

From manual orchestration experiments we have elicited a set of orchestration rules and devised a formal model for their implementation in an automated system. We have implemented such a system and evaluated it both off-line and in real-time. The preliminary results were promising. Our immediate research agenda includes a number of trials specifically designed for the evaluation of orchestrated communication, that are to be carried out with a more significant number of participants and quantitative evaluation methods, as we take these initial positive evaluations as sufficient to motivate further developments.

We have concluded that an expressive set of orchestration rules can be defined using a formalism based on the concept of channels of communication. We aim to extend this model further by using first-order logic to describe those rules. Moreover, we plan to enrich the current model with the ability to reason more generally with time. For that purpose, we will bring temporal constructs into first-order logic, allowing the system to reason about sequences of high-level events.

5. REFERENCES