Multiparty attention management for an embodied conversational agent

Léa Haefflinger, Frédéric Elisei, Gérard Bailly

Multiparty attention management for an embodied conversational agent

Léa Haefflinger[1,2], Frédéric Elisei[1] & Gérard Bailly[1]
[1]CRISSP, GIPSA-lab, Grenoble, France
[2]Innolab, Grenoble-France
{lea.haefflinger,frederic.elisei,gerard.bailly}@gipsa-lab.grenoble-inp.fr

I. Introduction

Humans converse via verbal and non-verbal cues. Non-verbal cues include gaze, head movements, facial expression, gestures, body position, etc. These cues complement speech with multiple information, such as who or what element of the physical scene is involved in the discourse (e.g. via deictic gestures or gaze). These cues are particularly important for multi-party conversations. Indeed, the higher the number of participants in the conversation, the more complex is the interaction. Participants can play other roles than just speaker and addressee, such as "side participant", "overhearer"... These roles shift during the conversation, called "Footing" [1]. These changes occur smoothly, as humans use implicit social codes, which can be verbal or non-verbal. A social robot should be able to detect and generate these codes. Mutlu and al [2] showed that a robot could inform the participants about their role by the sole use of gaze. Skantze and al [3] influence turn-taking and turn-holding and so impact who will be the next speaker, just by controlling the robot's attention. Gillet and al [4] used robot's gaze to balance participation in conversation. Moreover, gaze impacts engagement of participants in conversation [5], recall of informations [6], etc.

Providing a social robot with attention management is therefore a prequisite for monitoring role and information processing. Our ambition is to endow robots with such a skill using a data-driven approach: we first collect ground-truth behavioral data via an original immersive teleoperation platform where human pilots artificially endow the robot with such a skill. Then the multimodal behavioral scores are mined to extract behavioral models via machine learning techniques.

We first present our multiparty interactive scenario and the dataset we collected in the framework of the RoboTrio project. We will then present a detailed analysis of the head and eye mouvements of the pilot (driving the head and eyes of our robot that are independently controlled). We show that head and eye mouvements should be controlled independently and fullfill different functions, i.e. monitoring group vs. individual addressee.

II. Data collection

1) Dataset: The dataset comprises 22 sequences of human diads playing a game with our teleoperated robot iCub named Nina [7]. Nina reproduces the head, gaze and lips movements of a remote human pilot whose movements are tracked and streamed via real-time motion capture, in partiticular using a HMD equipped with an embedded binocular eyetracker. Each eye of the robot embedds a camera whose video stream is displayed in the corresponding field of the HMD. The pilot can thus see the two subjects facing the robot and a tablet in the hand of the robot where the game instructions are written.

For each sequence, the couple is different and is composed of two men or two women. Each interaction lasts between 17 and 25 minutes. the movements of all effectors (lips, eyes, head motion) and all sensors (stereo audio for ear microphones and video) of the robot are recorded. This endogenous data is complemented with audio from head microphones weared by interlocutors and videos of two fixed HD cameras.

The game is similar to Unanimo®: the two subjects have to guess the most quoted words related to a seed word, according to a survey (such as rose, sea, antennas... related to shrimp). The robot animates the game: it has to introduce the theme, encourage discussions, get consensus and give scores. Thanks to the use of teleoperation, the subjects have the illusion to interact with a skilled robot with a human-like behavior. This data collection framework nears training from inference conditions: interactions are already limited by sensorimotor capabilities of the robot and HRI a priori and expectations from human partners.

2) Data annotation: We already annotated completely 5 of the 22 sequences, all corresponding to men couple interactions. Several streams have been semi-automaticall labelled:

1) Robot's gaze: We automatically associate the robot's gaze to three regions of interest (RoI) via Gaussian mixtures: left vs. right subject, and the tablet. Occasional gaze aversion is labelled "elsewhere". After checking if the point is not corresponding to a saccade, fixations are allocated to these 4 RoI.
2) Robot's head: The orientation of the pilot's head is also classified, depending on whether the head is facing left, right, center (between the two subjects) or down (tablet position).
3) Subjects gaze: Sujects gaze is classified with GMM too, the different classes are "Robot", "OtherSubject" and "Elsewhere". To do so, OpenFace [8] runs on HD videos to detect head and eyes orientation and then compute gaze focal points.
4) Robot's speech: The speech contents and intentions are annotated manually for the pilot. We have defined 24

different intentions for the pilot, for example: ”Theme announce”, ”Ask Proposition”, ”Ask Validation”, ”Give Positive Scoring”, ”Give Nul Scoring”...

5) Subjects speech: 9 intentions for subjects have been defined : ”Proposition”, ”Positive Feedback”, ”Negative Feedback”...

III. Results

1) Analysis of gaze according to pilot's intent: When the pilot is speaking, he looks most of the time at the tablet. But the gaze's distribution depends on speech's intent. For example, when the utterance is labeled ”Theme announce” or ”Scoring”, the pilot looks almost all the time at the tablet, because he needs the information writen on it. But when the label corresponds to a question addressed to the two subjects, he looks both with an equal repartition. To model the gaze behavior of the robot, it is important to take into account what the robot says.

2) Analysis of gaze according to the speaker's intent: A naive hypothesis would state that when one of the subjects is speaking, the pilot look at her/him. But in our data, the pilot looks at the other subject one third of the time. Moreover, when the speaker is looking at the other subject, the pilot's gaze distribution is equally balanced between the two. In short, to model the gaze behavior of the robot, it is important to take into account who is the speaker and who is the addressee: the robot, the other subject or both.

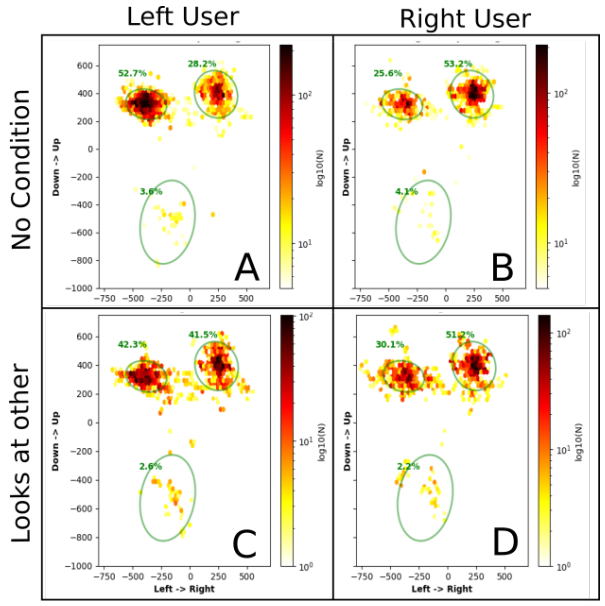

Fig. 1. Heat maps of pilot's focal points according to speaker, for one sequence. A/C The speaker is user left, B/D is user right. C/D The speaker looks at the other user. The percentages correspond to the proportion of points contained in the GMM ellipses representing the 3 targets.

3) Analysis of difference between head and gaze behavior: Both head and eye gaze focal points have been computed. Their distributions are different. Obviously, eye gaze focal points cover a larger area than head's ones. In several sequences, head points are mainly located between the two regions of interest corresponding to the left and right subjects. As put forward by Otuska and al [9], gaze cue the adressee while head direction is placed at the center of the cone of attention: the monitoring of the so-called ”conversational regimes” require an independent control of head and eye movements, where the former is not entirely slaved by the later.

IV. On going work & perspectives

Before training models, we will test what modelling parameters third-party overhears will prefer. To do so, we will present to raters small video clips where our robot replays its multimodal behaviour but with different control systems for the head and eye movements:

1) ground-truth: the robot replays the recorded behavior
2) eye-only: the robot only moves its eyes to gaze RoI
3) head-only: the robot only moves its head to gaze RoI
4) head-from-gaze: the head accompanies eye movements according to a scrutinization model [10]

Adresses, speech intents will then be used to predict gaze and head movements using predictive techniques such as tCNN. These control models will then be confronted to ground-truth and coupled with a real-time dialog system to perform conversation management. A style component will be added to cope with personnalities of human partners and actively monitor their turns.

References

[1] Goffman, Erving. 1979. ”Footing.” Semiotica 25 (1-2):1-30.

[2] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. ”Footing in human-robot conversations: how robots might shape participant roles using gaze cues.” 4th ACM/IEEE international conference on Human robot interaction.

[3] Gabriel Skantze, Martin Johansson, and Jonas Beskow. ”Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects.” 2015 ACM on International Conference on Multimodal Interaction.

[4] Sarah Gillet, Ronald Cumbal, André Pereira, José Lopes, Olov Engwall, and Iolanda Leite. ”Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels.” 2021 ACM/IEEE International Conference on Human-Robot Interaction.

[5] Yoshinori Kuno, Kazuhisa Sadazuka, Michie Kawashima, Keiichi Yamazaki, Akiko Yamazaki, and Hideaki Kuzuoka. 2007. ”Museum guide robot based on sociological interaction analysis.” SIGCHI Conference on Human Factors in Computing Systems.

[6] Bilge Mutlu, Jodi Forlizzi and Jessica K. Hodgins. 2006. “A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior.” 6th IEEE-RAS International Conference on Humanoid Robots 518-523.

[7] Metta, Giorgio Natale, Lorenzo Nori, Francesco Sandini, Giulio Vernon, D. Fadiga, Luciano Hofsten, Claes Rosander, Kerstin Lopes, Manuel Santos-Victor, José Bernardino, Alexandre Montesano, Luis. 2010. ”The iCub humanoid robot: An open-systems platform for research in cognitive development.” Neural networks : the official journal of the International Neural Network Society. 23. 1125-34.

[8] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. ”OpenFace 2.0: Facial Behavior Analysis Toolkit.” IEEE International Conference on Automatic Face and Gesture Recognition.

[9] KKazuhiro Otsuka. 2011. ”Conversation Scene Analysis [Social Sciences],” in IEEE Signal Processing Magazine, vol. 28, no. 4, pp. 127-131

[10] Laurent Itti, Nitin Dhavale and Frederic Pighin. 2006. ”Photorealistic Attention-Based Gaze Animation.” IEEE International Conference on Multimedia and Expo. pp. 521-524