Analyzing Gaze during Face-to-Face Interaction

Stephan Raidt, Gérard Bailly, Frédéric Elisei

Dept. of Speech & Cognition, GIPSA-Lab, Grenoble Universities – France
{Stephan.Raidt, Gerard.Bailly, Frederic.Elisei}@gipsa-lab.inpg.fr

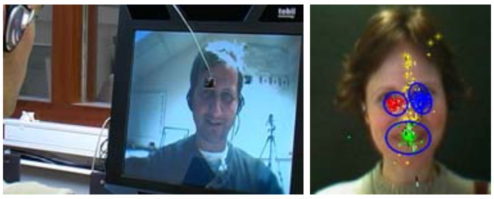
Abstract. We present here the analysis of multimodal data gathered during realistic face-to-face interaction of a target speaker with a number of interlocutors. Videos and gaze have been monitored with an experimental setup using coupled cameras and screens with integrated eye trackers. With the aim to understand the functions of gaze in social interaction and to develop a coherent gaze control model for our talking heads we investigate the influence of cognitive state and social role on the observed gaze behavior.

Keywords: face-to-face interaction, talking head, gaze control model.

Gaze is an essential component of face-to-face interaction. When interacting with a human interlocutor, the gaze patterns of an Embodied Conversational Agent (ECA) carry important cues not only for signaling the ECA's own communicative intentions but also its awareness of the environment as well as of the cognitive and emotional state of its interlocutors. The aim of our research is to understand the functions of gaze in social interaction and to develop a coherent gaze control model for our talking heads taking into account the results of detailed multimodal scene analysis.

Bilvi and Pelachaud [1] propose a gaze model for dyadic conversations. Textual input to the system augmented with tags indicating communicative functions drives a statistical model to generate eye movements alternating between direct and averted gaze. Lee et al [4] propose also a similar statistical model based on analysis of video recordings of monologues uttered by one subject. Both models take into account the cognitive activity of the ECA (e.g. speaking vs. listening, etc) but do not integrate any detailed scene analysis: they do not determine exactly where their target speakers are looking. Itti et al. [3] developed a gaze control model coupled with a visual attention system that detects salient and pertinent points of interest in a natural scene and triggers exogenous saccades. There is however no detection nor separate treatment of faces in this system.

We developed an experimental platform where two subjects can interact via a crossed camera–screen setup, with the aim to give interlocutors the impression to be facing each other across a table. Video and audio signals as well as gaze directions are recorded during the interaction for later analysis. The experiment involves dyadic interactions between a female reference subject (for whom we build a virtual speaking clone) and several naïve subjects with same social status and sex. The interaction consists in a sentence-repeating game. One partner utters semantically unpredictable sentences that the other is asked to repeat. Roles are further exchanged. With this rather restricted scenario we try to capture the main elements of face-to-face interaction and to enhance the gaze cues of mutual attention. It also imposes a clear chaining of cognitive states (reading, speaking, listening, waiting, thinking…) and simplifies complex negotiation of turn taking and state dependent gaze analysis. According to our knowledge this is the first experimental setup that monitors both subjects during such a mediated face-to-face interaction.



Left: Experimental setting. Right: distributions of fixations towards our reference speaker.

Based on the audiovisual and gaze data saccades and fixations of each speaker with reference to the respective position of the face of their interlocutor are automatically computed. Analysis of these data clearly confirms the triangular pattern of fixations scanning the eyes and the mouth previously obtained by Vatikiotis-Bateson, Eigsti et al [5] for perception of prerecorded audiovisual speech. The fixations towards four regions of interest (left and right eyes, mouth, and face) have been further distinguished and impact of role and cognitive state of each interlocutor are examined. We show that both factors have a significant impact on distributions of gaze fixations and blinking rate of our reference subject. We show for example that speakers never fixate the mouth of their interlocutors when speaking and that blink rate is accelerated when ‘speaking’, whereas ‘reading’ and ‘listening’ slow it down and often inhibit blinks.

Following general results obtained by Gullberg and Holmqvist [2] we also show that pre-recorded stimuli produce significantly different gaze behavior of the interlocutors compared to live interaction.

Based on the measured data we built a first gaze control model for our talking head by training a Hidden Markov Model. Given a succession of cognitive states with associated durations it computes parameters describing the fixations of the ECA towards the various regions on the face of its interlocutor. An initial state in each HMM has been added to cope with the particular distribution of the first fixation. The observation probabilities determine the duration of the fixation emitted by the HMM at each transition. They are computed from fixations gathered from the interactions. Fixations to the mouth are for instance longer than fixations to the eyes

References

[1] Bilvi, M. and C. Pelachaud. *Communicative and statistical eye gaze predictions*. in *International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 2003. Melbourne, Australia.

[2] Gullberg, M. and K. Holmqvist. *Visual attention towards gestures in face-to-face interaction vs on screen*. in *International Gesture Workshop*. 2001. London, UK.

[3] Itti, L., N. Dhavale, and F. Pighin. *Realistic avatar eye and head animation using a neurobiological model of visual attention*. in *SPIE 48th Annual International Symposium on Optical Science and Technology*. 2003. San Diego, CA.

[4] Lee, S.P., J.B. Badler, and N. Badler, *Eyes alive*. ACM Transaction on Graphics, 2002. **21**(3): p. 637-644.

[5] Vatikiotis-Bateson, E., et al., *Eye movement of perceivers during audiovisual speech perception*. Perception & Psychophysics, 1998. **60**: p. 926-940.