Embodied conversational agents : computing and rendering realistic gaze patterns

Gérard Bailly, Frédéric Elisei, Stephan Raidt, Alix Casari & Antoine Picot

Institut de la Communication Parlée, 46 av. Félix Viallet, 38031 Grenoble - France
{gerard.bailly, frederic.elisei, stephan.raidt, alix.casari, antoine.picot}@icp.inpg.fr

Abstract. We describe here our efforts for modeling multimodal signals exchanged by interlocutors when interacting face-to-face. This data is then used to control embodied conversational agents able to engage into a realistic face-to-face interaction with human partners. This paper focuses on the generation and rendering of realistic gaze patterns. The problems encountered and solutions proposed claim for a stronger coupling between research fields such as audiovisual signal processing, linguistics and psychosocial sciences for the sake of efficient and realistic human-computer interaction.

Keywords:. Embodied conversational agents, talking faces, audiovisual speech synthesis, face-to-face interaction.

1 Introduction

Building Embodied Conversational Agents (ECA) able to engage a convincing face-to-face conversation with a human partner is certainly one of the most challenging Turing test one can imagine (Cassell, Sullivan et al. 2000). The challenge is far more complex than the experimental conditions of the Loebner Prize[1] where dialog is conducted via textual information: the ECA should not only convince the human partner that the linguistic and paralinguistic contents of the generated answers to human inquiries have been built by a human intelligence, but also generate the proper multimodal signals that should fool human perception. We are however very close to being able to conduct such experiments. Automatic learning techniques that model perception/action loops at various levels of human-human interaction are surely key technologies for building convincing conversational agents. George, the talkative bot that won the Loebner Prize 2005, learned its conversation skills from the interactions it had with visitors to the Jabberwacky website, and through chats with its creator, Mr Carpenter. Similarly the first Turing test involving a non interactive virtual speaker (Geiger, Ezzat et al. 2003) has demonstrated that image-based facial animation techniques are able to generate and render convincing face and head movements.

[1] The Loebner Prize for artificial intelligence awards each year the computer program that delivers the most human-like responses to questions given by a panel of judges over a computer terminal.

Combining a pertinent dialog management with convincing videorealistic animation is still not sufficient to reach a real sense of presence (Riva, Davide et al. 2003). The sense of “being there” requires the featuring of basic components of situated face-to-face communication such as mixed initiative, back channeling, turn taking management, etc. The interaction requires a detailed scene analysis and a control loop that knows about the rules of social interaction: the analysis and comprehension of an embodied interaction is deeply grounded in our senses and actuators and we do have strong expectations on how dialogic information is encoded into multimodal signals.

Appropriate interaction loops have thus to be implemented. They have to synchronize at least two different perception/action loops. On the one hand there are low-frequency dialogic loops. They require analysis, comprehension and synthesis of dialog acts with time scales of the order of a few utterances. On the other hand there are interaction loops of higher frequency. These include the prompt reactions to exogenous stimuli such as sudden events arising in the environments or eye saccades of the interlocutor. The YTTM model (Thórisson 2002) of turn-taking possesses three layered feedback loops (reactive, process control and content). Content and reactive loops correspond to the two loops previously sketched. The intermediate process control loop is responsible for the willful control of the social interaction (starts and stops, breaks, back-channeling, etc). In all interaction models, information- and signal-driven interactions should then be coupled to guarantee efficiency, believability, trustfulness and user-friendliness of the information retrieval.

We describe here part of our efforts for designing virtual ECAs that are sensitive to the environment (virtual and real) in which they interact with human partners. We focus here on the control of eye gaze. We describe the multiple scientific and technological challenges we face, the solutions that have been proposed in the literature and the ones we have implemented and tested.

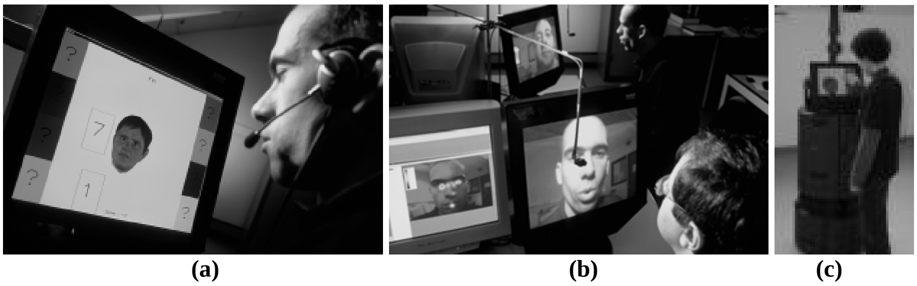


Figure 1: Face-to-face interaction: (a) gaming with an ECA; (b) studying human gaze patterns; (c) our ECA mounted on the Rackham mobile robot at the Space city in Toulouse – France (Clodic, Fleury et al. 2006). Copyright CNRS for (a) and (b).

2 Gaze and mutual gaze patterns.

The sampling process with which the eye explores the field of sight consists of fixations, smooth pursuits and saccades. Saccades are the rapid eye movements (approx. 25-40ms duration) with which the high-resolution central field (the fovea) is pointed to the area of interest. Fixations (and slow eye movements) of relatively long duration (300ms) enable the visual system to analyze that area (e.g. identify objects or humans). They are characterized by microsaccades that compensate for retinal adaptation. Functionally these two components correspond to two complementary visual streams, a ‘where’- and a ‘what’-stream (Grossberg 2003). The ‘what’-stream is responsible for object recognition, the ‘where’-stream localises where these objects and events are. The ’what’-stream is assumed to be allocentric, i.e. object centered, whereas the ‘where’-stream is egocentric, i.e. observer centered. An additional mechanism, called smooth pursuit, locks slowly moving interest points in the fovea.

Scrutinizing a scene (either a static picture or a video) is more complicated than just moving from one salient feature of the scene to the next. Perceptual salience is not the only determinant of interest. The cognitive demand of the scrutinizing task has a striking impact on the human audiovisual analysis of scenes and their interpretation. Yarbus (1967) showed notably that eye gaze patterns are influenced by the instructions given to the observer during the examination of pictures. Similarly Vatikiotis-Bateson et al (1998) showed that eye gaze patterns of perceivers during audiovisual speech perception are influenced both by environmental conditions (audio signal-to-noise ratio) and by the recognition task (identification of phonetic segments vs. the sentence’s modality). Attention is also essential. As an example the work of Simons and Chabris (1999) suggests that attention is essential to consciously perceive any aspect of a scene. Major changes to objects or scenes may be ignored ('change blindness') and objects may even not be perceived (‘attentional blindness’) if they are not in our focus of attention.

Finally, eye gaze is an essential component of face-to-face interaction. Eyes constitute a very special stimulus in a visual scene. Gaze and eye-contact are important cues for the development of social activity and speech acquisition (Carpenter and Tomasello 2000): theories of mind[2] (Scassellati 2001) rely on the ability of computing eye direction of others. In conversation, gaze is involved in the regulation of turn taking, accentuation and organization of discourse (Argyle and Cook 1976; Kendon 1967). We are also very sensitive to the gaze of others when directed towards objects of interest within our field of view or even outside (Pourtois, Sander et al. 2004). In the Posner cueing paradigm (1980), observers’ performance in detecting a target is typically quicker in trials in which the target is present at the location indicated by a former visual cue than in trials in which the target appears at the uncued location. The outstanding prominence of the human face in this respect was shown by Langton et al. (1999; 2000). Driver et al. (1999) have shown that a concomitant eye gaze also speeds reaction time.

The data presented so far show that gaze control is a complex cognitive activity that not only depends on the environment – that of course includes others – but also on our own cognitive demands.

[2] The ability to understand that others have beliefs, desires and intentions that are different from one's own (Baron-Cohen, Leslie et al. 1985; Premack and Woodruff 1978)

3 Computational models for the observation of natural scenes

Most robots incorporate a computational model for observing their environment. Mobile robots use the results for planning displacements and avoid obstacles. Anthropoid robots embed cameras at eyes location and the movements that are necessary for controlling their field of view informs indirectly human partners on their focus of interest. Most sociable anthropoid robots control gaze for communication needs: robots constructed by the Humanoid Robotics Group at the MIT Artificial Intelligence Laboratory have been designed to mimic the sensory and motor capabilities of the human system. The robots should be able to detect stimuli that humans find relevant, should be able to respond to stimuli in a human-like manner. The first computational theory of mind built by Scassellati (Scassellati 2001) was already incorporating a complex control of eye gaze et neck movements for pointing and signalling shared visual attention. Robita developed at Waseda University (Matsusaka, Tojo et al. 2003) points to objects and regulates turn taking in group conversation by gaze direction.

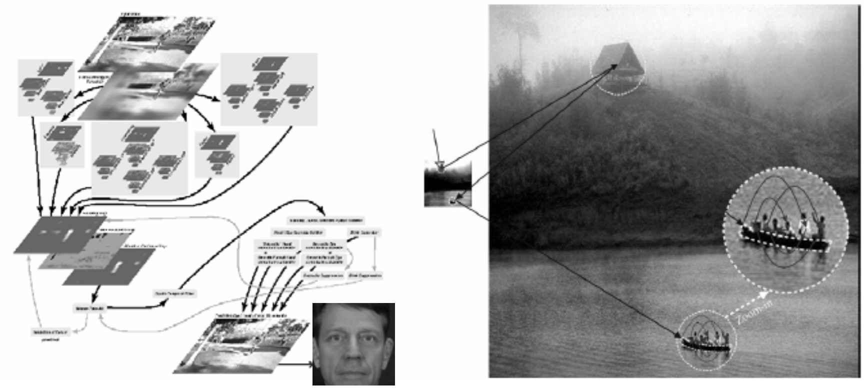


Figure 2: Models for observing natural scenes. Left: eye saccades of the ECA developed by Itty et al (Itti, Dhavale et al. 2003) are sequenced by points of interest computed from a video input. Right: Sun (Sun 2003) uses a multiscale segmentation to scrutinize an image by successive zoom-ins and -outs.

Most gaze control strategies for ECA are more elementary. When no contextual audiovisual stimuli are available (e.g. for web-based ECA), the basic strategy consists in globally reproducing blinks and gaze paths learnt by statistical models from human data (Lee, Badler et al. 2002). Attempts to regulate an ECA gaze from video input are quite recent: Itti et al (2003) propose a visual attention system that drives the eye gaze of an ECA from natural visual scenes. This system consists in computing three maps: (a) a saliency map, a bottom-up path that computes a global saliency for each pixel of the current image that combines color, orientation and flow cues; (b) a pertinence map, a top-down path that modulates the saliency map according to cognitive demands (e.g. follow white objects… that may cause attention blindness to events connected to darker areas of the scene), and (c) an attention map that is responsible with the observation strategy that switches between the successive points of interest.

The attention map also handles temporary Inhibition Of Return (IOR) so that all points of interest in a scene have a chance to be in focus. Although mostly tested on still images, the object-based attention framework proposed by Sun (2003) is based on a multi-scale segmentation of the image that computes a hierarchy of the points of interest as function of salience, granularity and size of the objects.

We recently implemented a eye gaze control system that builds on Itti et al proposal but replaces the pertinence and attention maps with a detector/tracker of regions of interest as well as with a temporary inhibition of return that rules the content of a small attention stack (Xu and Chun 2006) that memorizes position and appearance of previous regions of interest. The object detector is responsible for detecting known objects (such as faces) that triggers further predetermined hierarchical scrutation (such as focus on mouth and eyes for speaking faces) and for building statistical models of the shape and appearance of unknown objects (based yet on color histogram). If necessary, the detector uses the built characteristics to perform a smooth pursuit using a Kalman filter (see Figure 3). Once the object of interest does not move and fixation has been long enough for recognizing/building a model of the object, the object is pushed in the attention stack and the system seeks for the next salient object. While none is found, the system pops back the objects stored in the attention stack. The stack is also used for storing temporally the characteristics of an object that has not been entirely processed when a more salient object bumps in the scene: the exogenous stimulus is urgently processed and the system goes back to its normal sequential exploration.

Two natural videos have been used for testing (see Figure 3): the first scene features a subject waiving colored objects in front of him while the second one features several person passing behind a subject facing the camera. Gaze patterns computed by our system have been compared to human ones recorded using a non invasive Tobii® eyetracker: main differences occur when innate objects have a stronger intrinsic salience than faces in the scene (see Figure 4). Subjects are in fact more sensitive to faces than clothing since human faces are of most importance for understanding natural scenes. When interacting with people, events occurring in the immediate environment have also an impact on gaze and gaze interpretation. For instance, Pourtois et al (2004) have shown that facial expressions of your interlocutor is interpreted very differently depending on whether his gaze are directed to you or not.

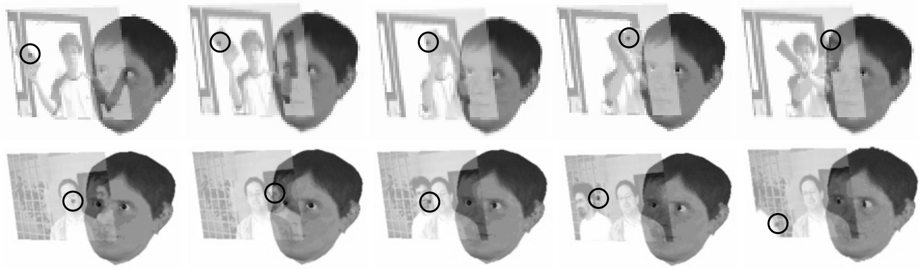


Figure 3: An ECA exploring a visual scene. The ECA scrutinizes a real scene displayed on a transparent screen. Key frames are displayed. A black circle materializes the point of interest for each image. Top: a subject waves a blue book in front of the ECA and the

module responsible for smooth pursuit controls the gaze. Bottom: a person passes behind the interlocutor and a saccade is performed to track this new object of interest.

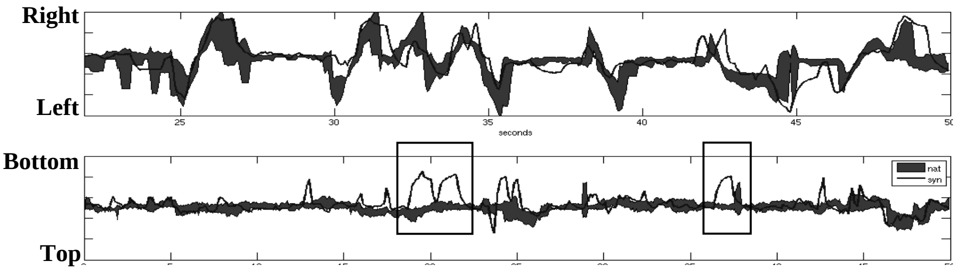
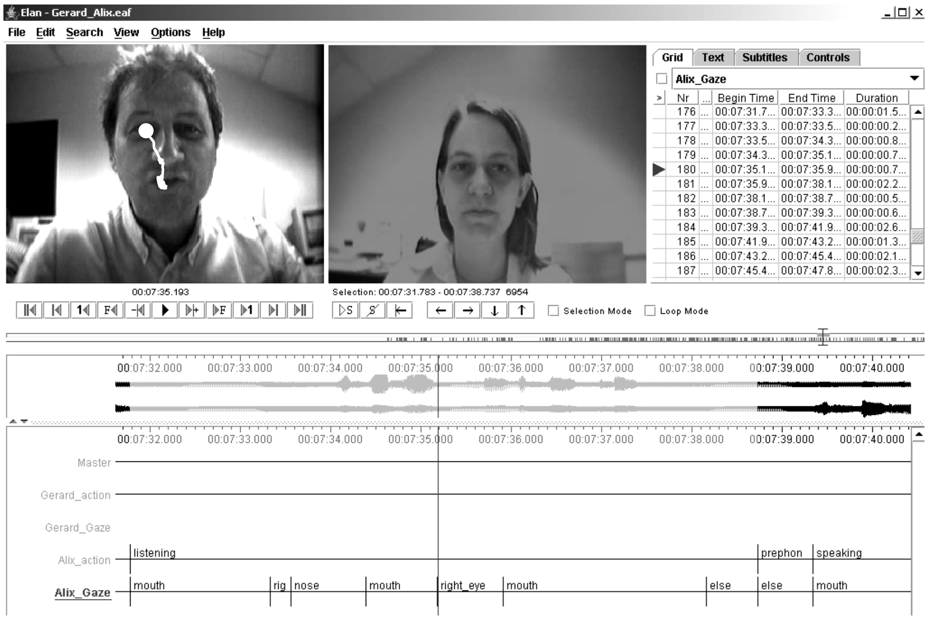


Figure 4: Comparing gaze trajectories (top: horizontal displacement; bottom vertical displacement) generated by our eye gaze control system with those recorded from subjects observing the same scene (the colored gauge is obtained by computing the variance between 5 subjects). Major differences (enlightened) are observed in vertical displacement where the control system is sometimes attracted by saturated colors of clothes of people passing in the background rather than their faces.



>	Nr	...	Begin Time	End Time	Duration
	176	...	00:07:31.7...	00:07:33.3...	00:00:01.5...
	177	...	00:07:33.3...	00:07:33.5...	00:00:00.2...
	178	...	00:07:33.5...	00:07:34.3...	00:00:00.8...
	179	...	00:07:34.3...	00:07:35.1...	00:00:00.7...
▶	180	...	00:07:35.1...	00:07:35.9...	00:00:00.7...
	181	...	00:07:35.9...	00:07:38.1...	00:00:02.2...
	182	...	00:07:38.1...	00:07:38.7...	00:00:00.5...
	183	...	00:07:38.7...	00:07:39.3...	00:00:00.6...
	184	...	00:07:39.3...	00:07:41.9...	00:00:02.6...
	185	...	00:07:41.9...	00:07:43.2...	00:00:01.3...
	186	...	00:07:43.2...	00:07:45.4...	00:00:02.1...
	187	...	00:07:45.4...	00:07:47.8...	00:00:02.3...

00:07:35.193

Selection: 00:07:31.783 - 00:07:38.737 6954

Figure 5: Screenshot of the labeling framework for face-to-face interaction data (using the ELAN editor® www.mpi.nl/tools/elan.html). The female listener fixates either the mouth or the right eye of the male speaker when he is uttering a SUS utterance (see text).

4 Gaze patterns in face-to-face interaction

When interacting, people mostly gaze at the other’s face and gesturing. While speech is clearly audiovisual (Stork and Hennecke 1996), facial expressions and gaze also inform us about the physical, emotional and mental state of the interlocutor. Together with gesturing, they participate in signaling discourse structure, ruling turn taking and maintaining mutual interest. Context-aware ECA should be reactive to gaze patterns of their interlocutors and implement these complex interaction rules (Thórisson 2002). Most data on eye movement of perceivers during audiovisual speech perception have been gathered using non interactive audiovisual recordings (Vatikiotis-Bateson, Eigsti et al. 1998). Several experiments have however shown that live gaze patterns are significantly different from screening (Gullberg and Holmqvist 2001): social rules have in fact a strong impact on communication when interacting face-to-face.

We conducted preliminary experiments for determining the natural gaze interplays between interlocutors according to their social status, their roles in the conversation and the dialog task. We illustrate below the complex gaze patterns already observed in a simple task such as repeating the other's utterance. The experimental setting involves two cameras coupled with two eye trackers (see Figure 1b) that monitor the gaze patterns of the interlocutors when interacting through two screens. We checked that this setting enables an acceptable spatial cognition so that each interlocutor correctly perceives what part of his face the other is looking at. The task just consisted in a speech game where Semantically Unpredictable Sentences (see Benoît, Grice et al. 1996, for description of SUS) uttered by one speaker in noisy environment have to be repeated with no error by his interlocutor. The speaker has of course to correct the repeated utterance as long as the repetition is incorrect. Mutual attention is thus essential to the success of interaction. Preliminary results (see Table 1) confirm for example that prephonatory (preparing to speak) activity is characterized by a gaze away from the face of the interlocutor. Eyes and mouth are all scrutinized when first listening to SUS whereas gaze during verification is focused on the mouth: gaze patterns are of course highly depending on cognitive demands (Yarbus 1967).

Table 1: Gaze data from speaker X when interacting with speaker Y. A turn consists in trying to repeat a SUS uttered by the partner with no error. Percentage of time spent on mouth and eyes regions is given for various actions and roles of the interlocutors.

		Regions of the face of Y gazed by X			
SUS giver	Actions of X	Mouth	Left eye	Right eye	Other
X	Prephonatory	48,1	0	6,9	45,0
	Speaking	91,6	0	5,1	3,3
	Listening	82,0	0	6,7	11,3
Y	Listening	64,0	14,6	17,8	3,6
	Speaking	48,4	29,2	19,2	3,2
	Prephonatory	18,7	10,2	37,6	33,5

5 Comments

A control model for eyes direction should not only rely on a context-aware multimodal scene analysis and a basic comprehension of the user's intentions and social rules but also rely on a faithful scene synthesis. Gaze patterns should be rendered so that human partners perceive the intended multimodal deixis and mutual attention. In a preceding paper (Raidt, Bailly et al. 2006), we have shown that our ECA is able to efficiently attract users' attention towards its focus of interest. We currently investigate the impact of the eye gaze rendering on performance. Eyelids deformations as well as head movements participate to the elaboration of gaze direction: adequate prediction of these deformations according to gaze direction reinforces perception of spatial cognition.

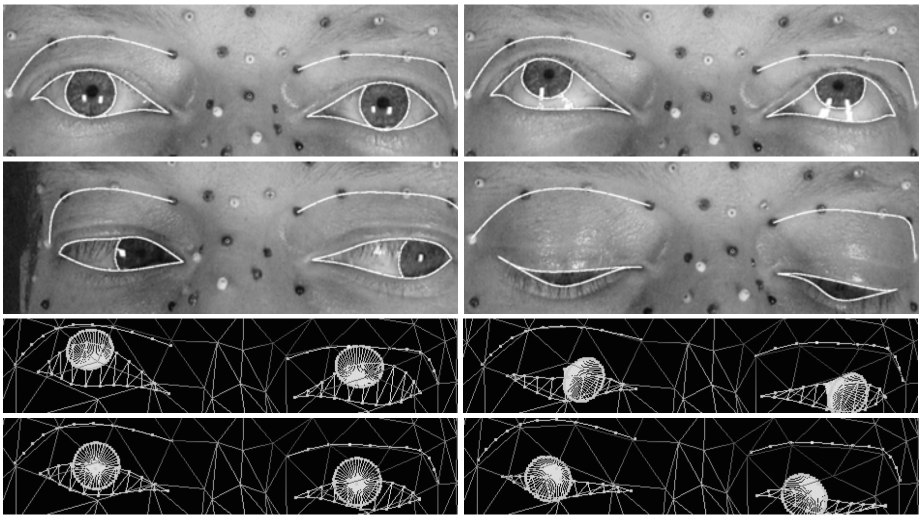


Figure 6: A 3D statistical shape model that reproduces geometric deformations of the eyelids of one subject depending on gaze direction.

6 Conclusions

This paper sketches a research framework for giving ECA the gift of situated human interaction. The landscape on eye gaze research is of course incomplete and gaze is one part of the facial actions that humans involve in face-to-face conversation. Gestural scores should be properly orchestrated so that complementary and redundant information is delivered at the right tempo to the interlocutor. Human behavior is so complex and subtle that computational models should be grounded on quantitative data (please refer for example to Bailly, Elisei et al. 2006, for a study of facial movements involved in conversation). Interaction rules should be completed with interaction loops that take into account the necessary coupling between signals extracted by a detailed multimodal scene analysis and the comprehension of the discourse and speaker's desires and beliefs that the artificial intelligence is able to built. Part of the success and realism of the interaction is surely in the intelligent use the artificial intelligence can make of the symptoms of the comprehension of the interaction the human partners who are present in the scene offer for free.

7 Acknowledgments

The authors want to thank Jan Cornelis, Hichem Sahli and Werner Verhelst for their kind invitation. The technical support from Alain Arnal and Christophe Savariaux was essential. This work is financed by the GIS PEGASUS, the ELESA research federation and the Presence project of the Cluster Rhones-Alpes InfoLog.

References

Argyle, M. and M. Cook (1976). Gaze and mutual gaze. London, Cambridge University Press.

Bailly, G., F. Elisei, P. Badin and C. Savariaux (2006). Degrees of freedom of facial movements in face-to-face conversational speech. International Workshop on Multimodal Corpora, Genoa - Italy: 33-36.

Baron-Cohen, S., A. Leslie and U. Frith (1985). “Does the autistic child have a “theory of mind”?” Cognition **21**: 37-46.

Benoît, C., M. Grice and V. Hazan (1996). “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences.” Speech Communication **18**: 381-392.

Carpenter, M. and M. Tomasello (2000). Joint attention, cultural learning and language acquisition: Implications for children with autism. Communicative and language intervention series. Autism spectrum disorders: A transactional perspective. A. M. Wetherby and B. M. Prizant. Baltimore, Paul H. Brooks Publishing. **9:** 30–54.

Cassell, J., J. Sullivan, S. Prevost and E. Churchill (2000). Embodied Conversational Agents. Cambridge, MIT Press.

Clodic, A., S. Fleury, R. Alami, R. Chatila, G. Bailly, L. Brèthes, M. Cottret, P. Danès, x. Dollat, F. Elisei, I. Ferrané and M. Herrb (2006). Rackham: an interactive robot-guide. IEEE International Workshop on Robots and Human Interactive Communications, Hatfield, UK

Driver, J., G. Davis, P. Riccardelli, P. Kidd, E. Maxwell and S. Baron-Cohen (1999). “Shared attention and the social brain : gaze perception triggers automatic visuospatial orienting in adults.” Visual Cognition **6** (5): 509-540.

Geiger, G., T. Ezzat and T. Poggio (2003). Perceptual evaluation of video-realistic speech. CBCL Paper #224/AI Memo #2003-003, Cambridge, MA, Massachusetts Institute of Technology.

Grossberg, S. (2003). “How does the cerebral cortex work? development, learning, attention, and 3d vision by laminar circuits of visual cortex.” Behavioral and Cognitive Neuroscience Reviews **2**: 47-76.

Gullberg, M. and K. Holmqvist (2001). Visual attention towards gestures in face-to-face interaction vs. on screen. International Gesture Workshop, London, UK: 206-214.

Itti, L., N. Dhavale and F. Pighin (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. SPIE 48th Annual International Symposium on Optical Science and Technology, San Diego, CA: 64-78.

Kendon, A. (1967). “Some functions of gaze-direction in social interaction.” Acta Psychologica **26**: 22-63.

Langton, S. and V. Bruce (1999). “Reflexive visual orienting in response to the social attention of others.” Visual Cognition **6** (5): 541-567.

Langton, S., J. Watt and V. Bruce (2000). “Do the eyes have it ? Cues to the direction of social attention.” Trends in Cognitive Sciences **4** (2): 50-59.

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

Matsusaka, Y., T. Tojo and T. Kobayashi (2003). “Conversation Robot Participating in Group Conversation.” IEICE Transaction of Information and System **E86-D** (1): 26-36.

Posner, M. I. (1980). “Orienting of attention.” Quarterly Journal of Experimental Psychology **32**: 3-25.

Pourtois, G., D. Sander, M. Andres, D. Grandjean, L. Reveret, E. Olivier and P. Vuilleumier (2004). “Dissociable roles of the human somatosensory and superior temporal cortices for processing social face signals.” European Journal of Neuroscience **20**: 3507-3515.

Premack, D. and G. Woodruff (1978). “Does the chimpanzee have a theory of mind?” Behavioral and brain sciences **1**: 515-526.

Raidt, S., G. Bailly and F. Elisei (2006). Does a virtual talking face generate proper multimodal cues to draw user's attention towards interest points? Language Ressources and Evaluation Conference (LREC), Genova - Italy: 2544-2549.

Riva, G., F. Davide and W. A. IJsselsteijn (2003). Being there: concepts, effects and measurements of user presence in synthetic environments. Amsterdam, IOS Press.

Scassellati, B. (2001). Foundations for a theory of mind for a humanoid robot. Department of Computer Science and Electrical Engineering. Boston - MA, MIT**:** 174 p.

Simons, D. J. and C. F. Chabris (1999). “Gorillas in our midst: sustained inattentional blindness for dynamic events.” Perception **28**: 1059-1074.

Stork, D. G. and M. E. Hennecke (1996). Speechreading by Humans and Machines. Berlin, Germany, Springer.

Sun, Y. (2003). Hierarchical object-based visual attention for machine vision. PhD Thesis. Institute of Perception, Action and Behaviour. School of Informatics. Edinburgh, University of Edinburgh**:** 169 p.

Thórisson, K. (2002). Natural turn-taking needs no manual: computational theory and model from perception to action. Multimodality in language and speech systems. B. Granström, D. House and I. Karlsson. Dordrecht, The Netherlands, Kluwer Academic**:** 173–207.

Vatikiotis-Bateson, E., I.-M. Eigsti, S. Yano and K. G. Munhall (1998). “Eye movement of perceivers during audiovisual speech perception.” Perception & Psychophysics **60**: 926-940.

Xu, Y. and M. M. Chun (2006). “Dissociable neural mechanisms supporting visual short-term memory for objects.” Nature **440**: 91-95.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. Eye Movements and Vision'. L. A. Riggs. New York, Plenum Press. **VII:** 171-196.