Multimodal face-to-face interaction with a talking face: mutual attention and deixis

Gérard Bailly, Frédéric Elisei, Stephan Raidt

Multimodal Face-to-Face Interaction with a Talking Face: Eye Gaze, Mutual Attention and Deixis

G. Bailly, F. Elisei & S. Raidt

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex, France
{bailly,elisei,raidt}@icp.inpg.fr

Abstract

Our long-term goal is to build an embodied conversational agent able to maintain realistic face-to-face communication with a human interlocutor. This conversational agent is embodied by a videorealistic talking head. While most researchers focus on discourse interpretation and generation, the main challenge here is to provide the interlocutor with implicit and explicit signs of mutual interest and attention as well as with an awareness of environmental conditions in which interaction takes place. A hybrid platform, with hardware and software, has been developed to test various interaction scenarios. As an application, the talking agent was used to interact with a user during a simple card game. The role of the agent was to act as a guide as well as to provide different levels of guidance (with or without mutual attention, and with or without endogenous eye saccades toward a correct or incorrect play). We provide here a comparative analysis of user performance across different levels of guidance and user perception of both level of guidance and level of help from the embodied conversational agent.

1 Introduction

Two complementary perspectives coexist implicitly in the development of Embodied Conversational Agents (ECA). The dialogic perspective (Cassell, Sullivan et al. 2000) focuses on the study of communicative interaction, with strong semantic and linguistic components, between human and/or software agents in mediated information systems. This perspective considers that the ultimate goal of interaction is information retrieval with ECA being the communication interface. The sociable perspective (Brooks, Breazeal et al. 1999; Breazeal 2002) puts forward the embodiment. In this later perspective our analysis and comprehension of an interaction is deeply grounded in our senses and actuators and we do have strong expectations on how dialogic information – if any - is encoded into multimodal signals. Of course users' mental representations and states, common belief spaces built when interacting with ECA is a complex construct that takes into account both communicative and sociable dimensions of interaction. Appropriate interaction loops have to be implemented. They have to synchronize low-frequency dialogic loops - that require analysis, comprehension and synthesis of dialog acts with time-scales of the order of a few utterances - with more high-frequency interaction loops - that require prompt reactions to the scene analysis such as involved in eye contact or exogenous saccades. Both information- and signal-driven interactions should be then coupled to guarantee efficiency, believability, trustfulness and user-friendliness of the information retrieval.
The work described here is dedicated to the analysis, modelling and control of multimodal face-to-face interaction between an embodied virtual conversational agent and a user. We particularly study here the impact of mutual attention in a series of simple deictic tasks.

1.1 Dimensions of face-to-face interaction

Building an ECA that may engage into a face-to-face interaction/conversation with a human partner is quite challenging. Not only the ECA has to decode the user's needs and intentions through multimodal communication, but also must give direct and indirect signs that it actually knows about *where* the interaction is taking place, *who* is its interlocutor and *what* ambient/localized service it may provide to the user(s). Such a rich face-to-face interaction (see Figure 1) requires intensive collaboration between the scene analysis and the specification of the task to be performed in order to generate appropriate actions of the ECA. Scene analysis performed by human beings is known to be permeable to cognitive demands. Yarbus (1967) for example instructed a subject to answer seven different questions about the depicted situation in Repin's picture "An Unexpected Visitor". Resulting eye gaze patterns show

that eyes tend to be attracted by those parts of the scene that contain the most information for the perception of it; not even salient visual features – contours, spotlights… - are given much attention, unless they convey important information for the recognition of the scene. Similarly Vatikiotis-Bateson et al (1998) show 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).

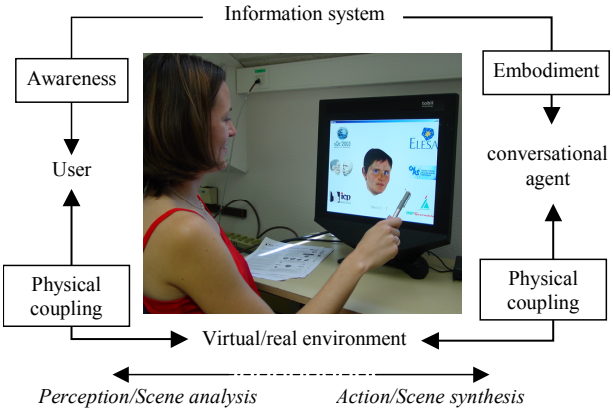

Figure 1: Embodied conversational agents and ambient interaction.
The control of agent actions should be aware of the user, the environmental conditions of the interaction and the competence of the information system to provide the user with relevant and reliable information. This involves a strong coupling between scene analysis and synthesis.

These complex eye gaze patterns constitute one of the dimensions of human activity that enables human beings to correctly attribute beliefs, goals, and percepts to other people. The set of abilities that allow an individual to infer these hidden mental states based on observed actions and behaviour is called a "theory of mind" (Premack and Woodruff 1978). Several TOM have been proposed (Baron-Cohen, Leslie et al. 1985; Leslie 1994). Baron-Cohen proposes for example an Eye Direction Detector (EDD) and an Intentionality Detector (ID) as basic components of a Shared Attention Mechanism (SAM) that is essential to the TOM's bootstrap. Actual implementation of these modules require the coordination of a large number of perceptual, sensorimotor, attentional, and cognitive processes. Scassellati (2001) developed an *embodied theory of mind* to link high-level cognitive skills to the low-level motor and perceptual abilities of a humanoid robot. The low-level motor abilities comprised coordinated eye, head and arm movements for pointing. The low-level perceptual abilities comprised essentially detection of salient textures and motion for monitoring pointing and visual attention.
We see here that even the unique control of eye gaze in face-to-face interaction is very complex and requires the coordination and cooperation of multiple processes, some being more particularly dedicated to the analysis of the multimodal scene whereas some others are more particularly concerned with interpreting the communicative intentions of the user that the information system may respond to (see Figure 1).

1.2 Eye gaze and attention

Some work suggests that attention may be required to consciously perceive any aspect of a scene. A drastic evidence for inattention blindness is provided by the work of Simons and Chabris (1999): without attention, we often do not detect large changes to objects and scenes ('change blindness') and may not even perceive objects. Visual attention can however be primed indirectly towards a certain area of the visual field using visual cues. In the Posner cueing paradigm (1990; 1980), observers' performance in detecting a target is typically better 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. Langton et al. (1999; 2000) have shown that observers react more quickly when the cue is an oriented face than an arrow. Driver et al. (1999) have shown that a concomitant eye gaze alone also speeds the reaction time.
Note that the amplitude of the benefit is quite small (20ms) compared to the total reaction time (around 300ms) and that observers are more sensitive to up/down visual cues than left/right ones. In the Posner cueing paradigm, the prime is however static and is presented very shortly (50ms) with a maximal effect when the prime is presented 300ms before the target.
The work presented below extends these studies towards practical use of eye saccades as cues to drive the direction of social attention.

1.3 Research aims

Our perspective is to develop an embodied TOM to link high-level cognitive skills to the low-level motor and perceptual abilities of a virtual conversational agent and to demonstrate that such a TOM will provide the information system with enhanced user satisfaction, efficient and robust interaction. The motor abilities is principally extended towards speech communication i.e. adapting content and speech style to pragmatic needs (e.g. confidentiality), speaker (notably age and possible communication handicaps) and environmental conditions (e.g. noise). If the use of a virtual talking head instead of a humanoid robot limits physical actions, it extends the domain of interaction to the virtual world: the user can also interact with other virtual objects (e.g. virtual icons) surrounding the virtual talking head (see the face-to-face system described below).

2 Developing our face-to-face platform

The user sits in front of a standard-looking flat panel screen, where a 3D talking head faces him or her, as shown on Figure 2. Hardware and software specificities allow the user to interact with the system using eye gaze, a mouse and speech. The 3D clone can look at the user, talk to him, and react to where the user looks. These elements form the basis of a grounded virtual face-to-face situation.

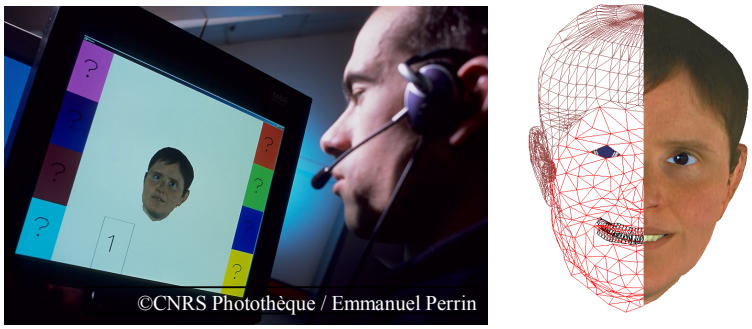

Figure 2: Face-to-face interaction (speech, gaze and mutual attention) with a 3D clone.

2.1 The hardware

The flat screen discretely embeds infrared lights and a camera. The Tobii1750 eye-tracker[1] allows us to analyse, at up to 60Hz, the eye gaze of the user whose head can move and rotate freely in a fairly unrestricted 3D volume (square cube of 40cm centred at 50cm away from the screen centre). Effective accuracy is obtained through a single short calibration procedure that each user must follow. Standard graphic hardware with 3D acceleration allows real-time rendering of the talking head on the screen.
A microphone/earphone headset allows oral communication between the two participants (the user and the agent) without audio interference.

2.2 Multimedia scenario scripting architecture

An event-based language has been developed, with a corresponding C++ code generator. This language allows the easy description and modification of multimedia scenarios. A finite-state automaton (FSA) describes each scenario as a series of states with pre-conditions and post-actions.
As the user progresses in the scenario, the FSA specifies which states are waiting for events. Pre-conditions consist in conjunctions or successions of expected multimodal events. Each event (e.g., recognized keywords, mouse clicks or displacements, eye movements, gaze to active objects) is time-stamped.
Pre-conditions can include tests on time-stamps intervals between events: this allows, for example, speech items (words that are identified as the sub products of the recognition of the linguistic content of an utterance) to be associated with a certain focus of attention.

[1] Please consult http://www.tobii.se/ for technical details.

Post-actions typically consist in the generation of multimodal events. Time-stamps of these events can be used to delay their actual instantiation in the future. Post-actions can also generate phantom events (to simulate multimodal input or to share information) that will be considered as potential triggers of pre-conditions of the states of the FSA.

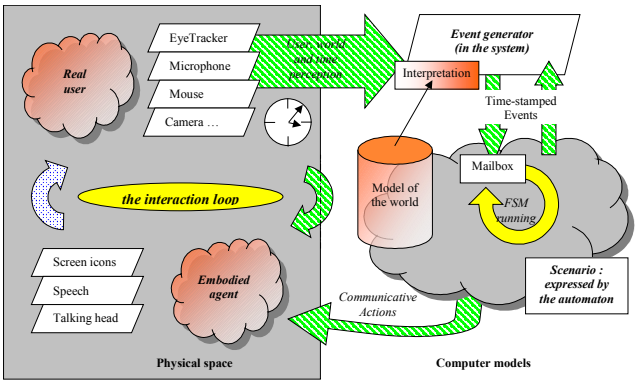

Figure 3: A finite state machine controlling the interaction.

Each input device (mouse, speech recognizer, eyetracker…) emits events according to users's actions and an internal model of the space of interaction that is refreshed constantly (see Figure 3). Triggerable area on the screen (selectable icons or parts of the talking head for example) are defined and tracked. Each time the user gaze points there, the system posts new events (a "zone entering" event, possibly following a "quitting previous zone" event) and potentially emits additional events such as "zone fixation duration" events. The FSA is called each time an event is generated or updated.
For accurate recording of the involved events and timings, the script is not interpreted but used to generate some C++ sources, which compile in a binary executable. The benefits of using C++ (variables, procedural and complex algorithms) remain accessible through code inclusion inside any script.

2.3 The talking head

We have cloned the 3D appearance and articulation gestures of a real human (Bailly, Bérar et al. 2003; Revéret, Bailly et al. 2000). The eye gaze of the clone can be controlled independently to look at the user, to look at where the user is looking on the screen (giving signs of mutual attention) or to direct the user's attention to 2D objects on the screen (vergence of the eyes is handled and provides a crucial cue for inferring spatial cognition). The virtual neck is also articulated and can accompany the eye-gaze movements.
The audiovisual messages can be either recorded by the original human speaker, or synthesized from text input. In practice the synthetic signals are here generated off-line to avoid slight reaction delays.

3 A multi-modal face-to-face interaction experiment

3.1 Our first scenario

We chose a card-playing scenario, where the user's task is to form pairs of cards bearing the same number. For this study, the user was asked to select cards using mouse clicks (the vocal recognition module was not used). After the user clicks on the first card, eight other cards are revealed simultaneously. The goal of the game is to click the right matching card as fast as possible (see Figure 4).

Four experimental conditions are considered: in the first one, no clone is displayed (as in Figure 4). In the second condition, the 3D head is visible and gives a random eye-saccade (with a partial head turn) to one of the non-matching cards just after the eight numbers are revealed (see left part of Figure 5). In the third condition, the saccade of the 3D head indicates the card to be matched (centre part of Figure 5). In the fourth condition, the cards are no longer revealed (the values cannot be read, question marks remain displayed, as shown on the right part of Figure 5), while the saccade remains the correct one: so, the saccade is the only available hint. In all the conditions, the oral messages are exactly the same (although the talking head is not visible in the first case): an invitation to find the

matching card, and random congratulations when it is found. In all the conditions, the eight card values are shuffled before each turn, so that colour associations or memory are of no use for the matching task.

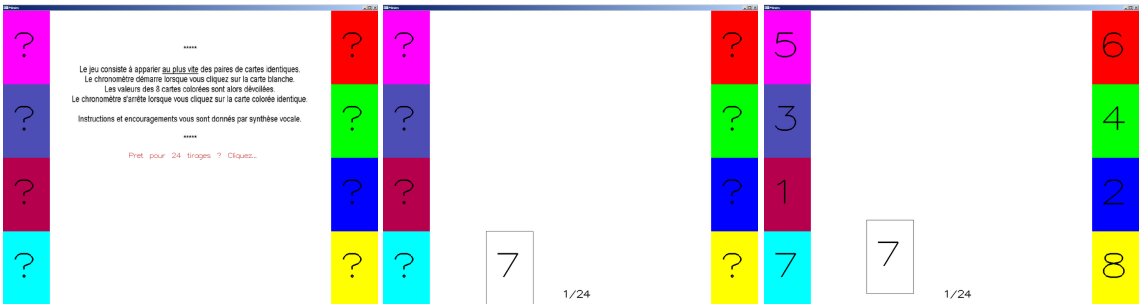

Figure 4: Experimental condition 1, where no clone is visible.
Left screen: The consigns are displayed, the values of the eight cards on the sides of the screen are hidden.
Middle screen: When the user clicks on the white card (here, the 7), the eight cards will be revealed.
Right screen: The card is now moving with the mouse, and the user must click the matching card (bottom left)

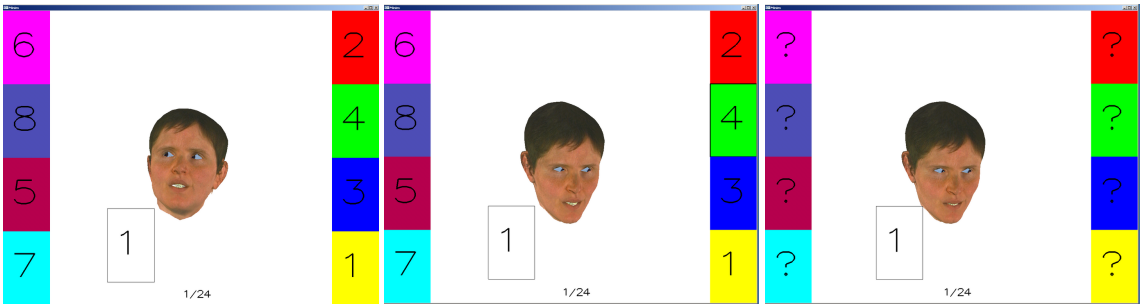

Figure 5: Comparative view of experimental conditions 2, 3 and 4 respectively, where the clone is visible. The user might be distracted *(first screen)* or helped *(second and third screen)* by an eye-saccade from the 3D clone. Please note that in condition 4 *(last screen),* cards were not revealed.

3.2 Data collection

Ten users (six male and four female) took part in the experiment, by playing the card-game described above. Participants ranged in age from 23 to 33 years, and most were students. All regularly used a computer mouse and none reported vision problems. The dominant eye was detected to be the right eye for all but one subject.

Each user had to play the game with the four successive experimental conditions described and illustrated above. Each experimental condition was initiated by a specific screen giving written instructions and followed by a corresponding training session with three card pairs to match. Through the written instructions, the user was aware when the clone would never (or respectively, always) look at the correct card, or that the card values would remain hidden. No strategy was suggested, but the user was instructed to find the matching pair as fast as possible. He or she could pause between two pairs and had 24 pairs to match. Each set of 24 looked random, but the eight possible final positions were addressed exactly three times.

During the four successive playing sessions, the time to match the right pair was measured (time from the click on the "dealt" card to time from the click on its match). We also recorded which cards the user looked at, and whether he or she looked at the clone (when visible), and if so, for how long. The eye-gaze activity (trajectories and positions within the active objects) was also tracked and recorded in a file for later analysis. A system log of all the posted messages was also recorded.

After the experiment (which lasted less than 20 minutes), participants filled out a form about various subjective aspects of the experiment, quantifying the following points:

- clone quality, ranked on a five-point scale: did the neck and eye movements look realistic? Was the eye gaze accurate enough to find the hidden cards in condition 4?

- task: did they think they were fast in some conditions? Were they influenced by the distracting eye gazes (condition 2)? Did they use the clone when it gave the correct solutions (conditions 3 and 4)?
- Experimental condition: Which on did they prefer (no clone, clone looking at the wrong place, clone looking at the right place, or hidden cards)? Which one would they choose to perform the task in the shortest amount of time possible?

3.3 Analysis of the experiment

3.3.1 Was the task achieved?

The game task allowed the user to place cards in wrong positions. Such mistakes could be identified from the recorded log files. Excluding training data, this never occurred in conditions 1 and 2. It occurred only once in condition 3 (clone looking at the correct card), and 34 times in condition 4 (cards remaining hidden). For this last condition, this amounted to a quite high error rate, 15% (34/240). All but one user made at least two mistakes (see first bar-plot in Figure 6).

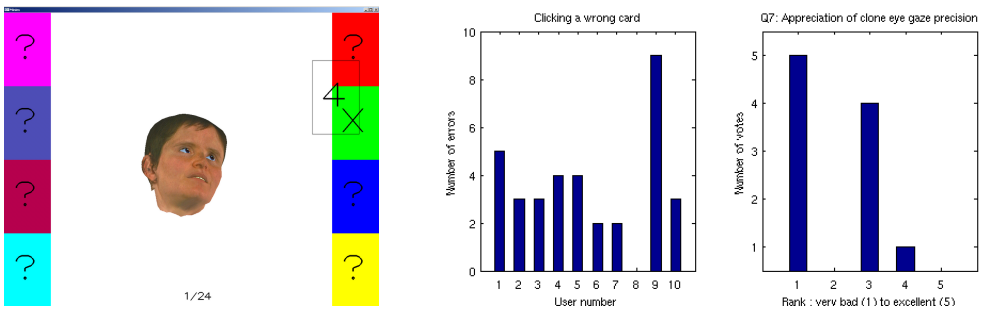

Figure 6: User clicking on an incorrect card: In the example on left, the clone was looking at the top right card, but the user clicked on the card below. Plots show that clicking a neighbouring card when clone eye gaze is the only available modality is a frequent error (only in condition 4, where users become aware of this).

The task being successfully performed in the first three experimental conditions, we will analyse the performance times to investigate whether the presence of a clone directing its eye gazes to the correct card or to other cards, improves or impairs user time efficiency.

3.3.2 Did the clone influence the performance?

The measured data is presented in three figures. Figure 7 illustrates the distribution of the time needed by each user to perform the task. With a similar presentation, Figure 8 details the time spent looking at a screen region of particular interest: this region corresponds to the clone mouth and eyes area and was also monitored for reference when no clone was visible. With the boxplots used in these two figures, we can observe the repartition of the data, and read the median of the measured times, which should be more robust than the mean to some outliers (represented by red crosses). Figure 9 illustrates with histograms the number of cards that were looked at by each user, including the correct one, before clicking on the right solution. These figures separate the results obtained in the conditions 1, 2 and 3. The time spent looking at cards was almost the complement of the time spent looking at the clone.

As expected with non-reflex complex tasks, some users are noticeably faster (e.g. user 10) or slower (e.g. user 1) than the majority, so direct comparison of answer times could not be performed.

Little time was spent looking at the clone area when the clone was not visible: user gazes were crossing this area rapidly while dragging the card or traversing the screen diagonally. In contrast, in condition 2 or 3, some users (but not all) have spent some time looking at the clone.

For condition 1 and 2, the number of cards looked at is rather high. This is obvious from the fact that looking at a card is the only way to find the searched one. For the third condition, the results in the bottom row show a different tendency for at least three users (3, 1 and 6).

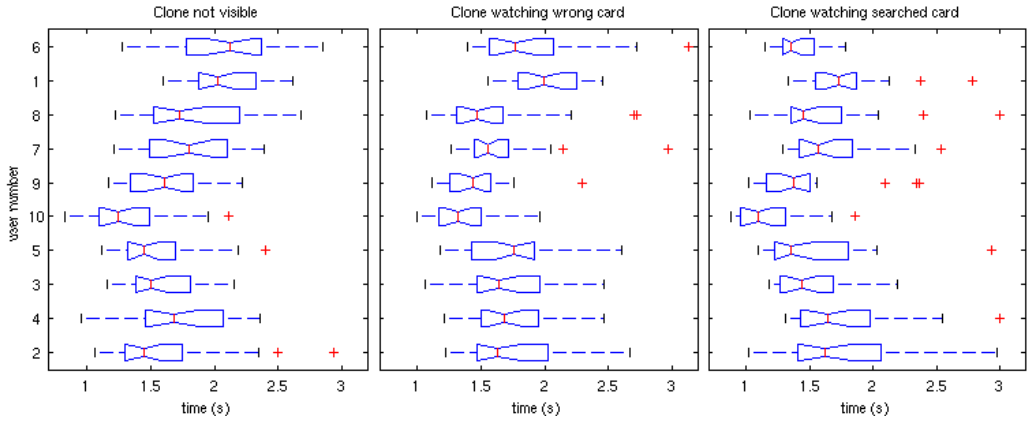

Figure 7: Time needed by the ten users to find the correct cards in conditions 1, 2 and 3 (box plots with median, two quartiles and outliers).

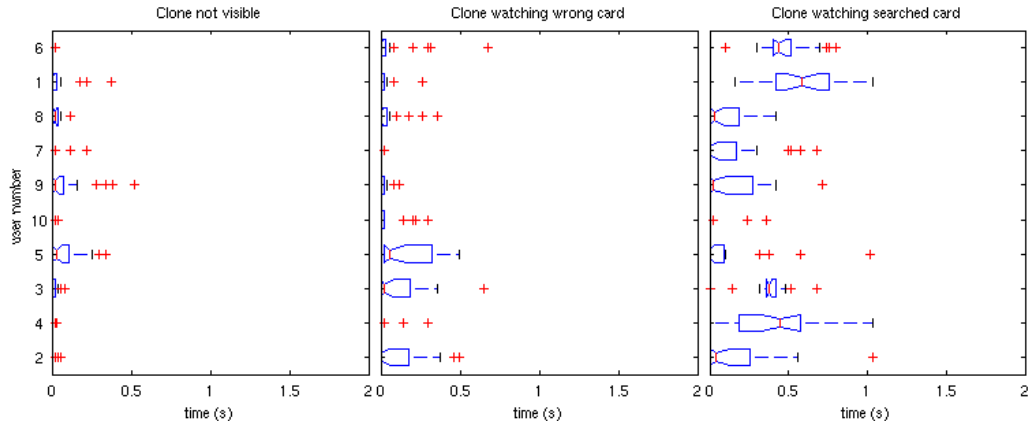

Figure 8: Time spent looking in the central region while seeking the correct card (whether or not the clone is present in this region).

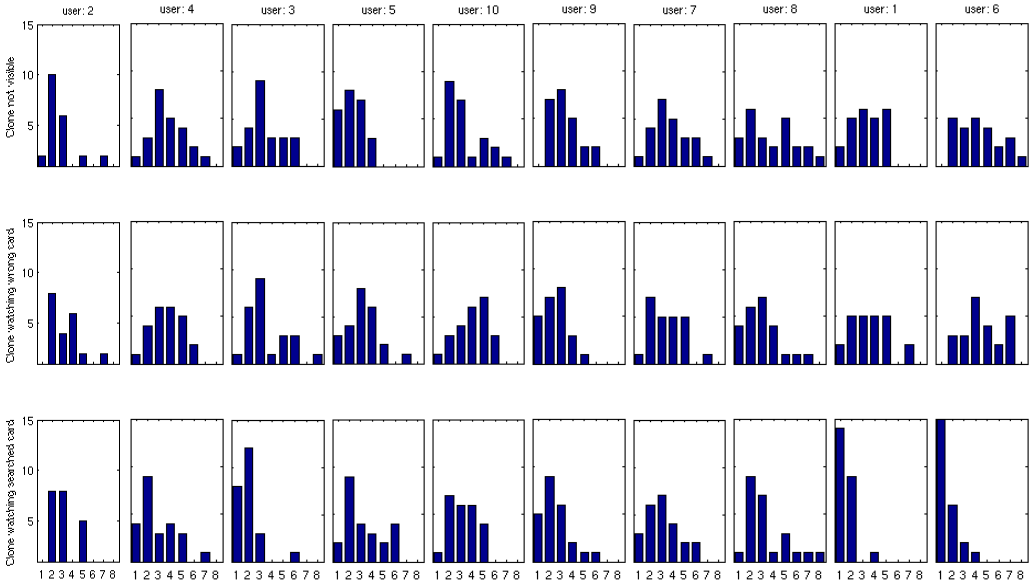

Figure 9: Histograms of the number of cards looked at by the ten users (columns) in the three conditions (rows)

Looking at these three figures, there is no obvious influence of the conditions on the measured results that would apply to all the users. Users seem to have different behaviours: clusters that may be found in one subplot cannot be found in another subplot. For example:

- user 4 spent almost the same time (around 1,7 seconds for the median) to perform the task, for the three conditions. In contrast, most users have significantly different time scores in the three conditions, as for example user 6 (respectively 2.1, 1.8 and 1.4 seconds for the median).
- users 7 and 9 did rarely look at the clone in condition 2, whereas users 5 and 2 looked at it enough to get a visible quartile.
- in condition 3 (clone directs its gaze to the correct solution), users 1 and 4 look durably at the clone (with half their gazes lasting more than 0.5 second), while user 10 nearly never looked at it.
- in condition 3, all but users 1, 3 and 6 looked at many cards before finding the correct one. With the help of the clone, user 1 could find the correct card after looking at one or two cards.

To help finding some behaviour groupings, we may compare the median of the time needed to perform the task in one condition with that needed in either of the other two conditions. There are three possible combinations, which might be used to sort the users: these three user orderings were used to obtain the plots in Figure 10. Interpretations based only on the median must of course be validated on the full data set.

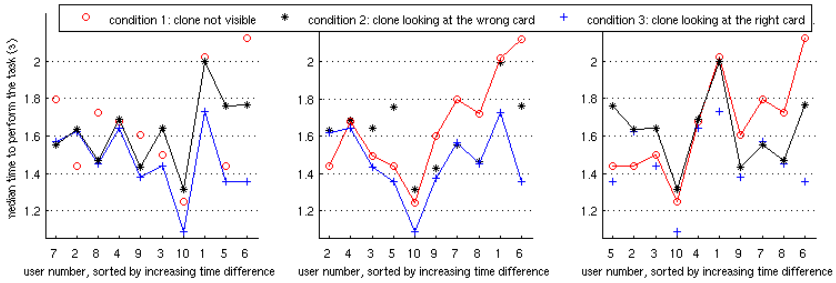

Figure 10: Efficiency (median time) to find the correct card, varying user sorting according to time differences

From the first subplot of this presentation, comparing median time of each user across condition 2 and 3, we note that users never performed slower in condition 3 than in condition 2. Except for users 4, 3, 1 and 6 that did look at the clone in condition 3, or users 5,3 and 2 that did look at the clone in condition 2, we should not conclude that the clone influenced the users. Possibly, habituation may explain some of the gains in the realisation of the task.
Indeed (third subplot), users 5 and 3, who were looking at the bad hints from the clone, were slower than in condition 1. We cannot tell if they choose to look at the clone (knowing the hints would be unproductive) or if they could not help looking at it.
All the users but user 2 performed faster (or within the same delay) in the condition 3 than in condition 1 (second subplot). Only users 3, 1 and 6 might have profited from the useful hints given by the eye gazes of the clone: they are the only users who invested significant time looking at the clone. Other users speed gains may, for example, be caused by habituation.

3.3.3 Did the users trust or use the clone?

For users 3, 1 and 6, we can confirm that they used the useful hints from the clone in situation 3: the number of cards they looked at is significantly lower: they could find respectively 23, 21 and 20 pairs (of the 24) while looking only at one or two cards. This proves that the 3D clone can be useful if users accept its guidance, although it unfortunately appears that the task we chose could be performed fast without the help of the clone, by brute force, looking at as many cards as necessary (users 10 or 4) and avoiding direct eye contact with the clone (the eye-tracker cannot record peripheral vision). This may explain why, in the post-experiment questionnaire, responses were not significantly different for the judgement on their speed (left graphs of Figure 11). Another conclusion regarding the influence of the clone is that users can escape its eye gaze. It is not salient enough - in our experimental conditions at least - to capture the attention of all the users. Indeed (right graphs of Figure 11), concerned users were conscious of the influence of some of the bad hints (user 5), or the good ones (users 3,5,6 and 1).

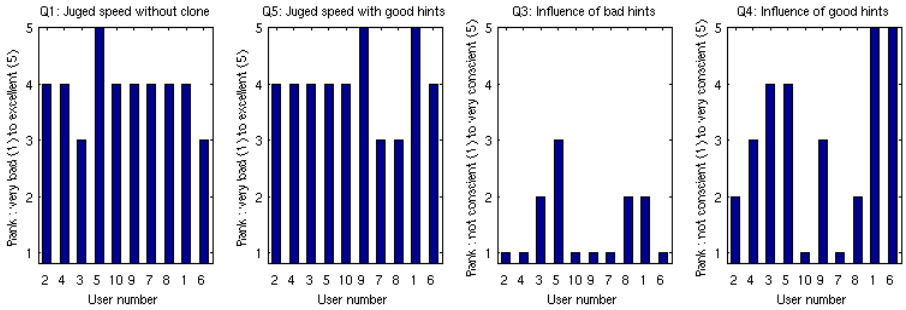

Figure 11: Impressions of the users about their speed and the influence of the eye gaze hints (in condition 1 and 3, and respectively in condition 2 and 3)

3.3.4 Discussion and perspectives

We described here our first effort at designing a system enabling face-to-face interaction between a 3D talking head capable of mutual attention and shared reality.
Of course, 3D rendering on a screen lacks the depth sensation and fails to transmit faithfully all the directions in the 3D world. Users ranked the use of this modality alone accordingly (condition 4, in the second bar-plot in Figure 6). It remains to be tested whether the neck and eye gaze animations could be improved to get better results (although the users ranked their realism as satisfactory, with five 'medium' marks and four 'good' marks). For this objective, we might use our interaction system to record eye gaze of real humans in the symmetrical situation (they would play the role of the trustworthy guide).
Despite the limitations of a display without stereovision and the lack of spatial accuracy in interpreting the clone eye gaze, our 3D clone could be used as an extra modality without causing selection errors during the clicking task.

Our first experimental scenario leaves a lot of freedom to the users, who used it to develop different strategies. We addressed an application that resembles our daily use of computer and mouse for information retrieval on screen, e.g. through file directories or icons: with this interesting choice, we collected representative data on the real interaction of our participants with the computer. Validating the interest of our platform in human computer interaction, we found that various strategies were used to perform the task. It might be interesting to explore in which case the cognitive load is higher. One way to achieve this would be to remove the possible pause between two successive cards, and to instruct users that they must act as fast as possible, without pausing. The number of cards could be increased, or cards could be revealed only when looked at. To check for habituation or fatigue, extra turns might be added, so that conditions might be repeated in various orders for the same participant.

4 Conclusions

We have developed a rather flexible hybrid platform (hardware and software system) that allows us to place users in a multi-modal face-to-face interaction loop with our talking agent and to record their activity for statistical analysis.
A first experiment was conducted with a playing card scenario creating a "too much information at the same time" situation, where an agent was proposed to help retrieve the correct information. We expected that using eye gazes of the 3D clone as an extra modality might lead to faster performances or lower the cognitive load. Preliminary analysis showed that users willing to use this level of guidance could perform the task faster or more easily: they could trust the clone and visit fewer cards. It also showed that some other users, choosing to ignore these hints whether they corresponded to incorrect solutions or correct ones, could succeed in this task, at least for the short time periods involved in our game.
We demonstrate that cues of mutual attention may benefit the performance in information retrieval. We believe that the study and modelling of the components of human face-to-face interaction are crucial elements of intuitive, robust and reliable communication. We are currently investigating interactive real-time eye-gaze patterns of human speakers in face-to-face communication with a special focus on the speaking/listening state. While most experimental data on speech and gaze examine attention of the listener, almost no experimental data is currently available on gaze patterns when speaking (Vertegaal, Slagter, van der Veer and Nijholt, 2001).

Acknowledgements

We gratefully acknowledge the patience of Hélène Lœvenbruck, the human model for our clone. We thank Alain Arnal and Christophe Savariaux for their technical assistance with the audiovisual capture platform, as well as Matthias Odisio, Pauline Welby and the reviewers for their helpful comments. We are also grateful to our experiment participants. This work would not have been possible with the results from several past projects involving students.

References

Bailly, G., M. Bérar, F. Elisei and M. Odisio (2003). "Audiovisual speech synthesis." International Journal of Speech Technology **6**: 331-346.

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

Breazeal, C. (2002). Designing Sociable Robots, The MIT Press.

Brooks, R. A., C. Breazeal, M. Marjanovic, B. Scassellati and M. Williamson (1999). The Cog Project: Building a Humanoid Robot" in Computation for Metaphors, Analogy, and Agents. Lecture Notes in Artificial Intelligence. C. Nehaniv. New York, Springer**:** 52–87.

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

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.

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.

Leslie, A. M. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. Mapping the Mind: Domain specificity in cognition and culture. L. A. Hirschfeld and S. A. Gelman. Cambridge, Cambridge University Press**:** 119–148.

Posner, M. and S. Peterson (1990). "The attention system of the human brain." Annual Review of Neuroscience **13**: 25-42.

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

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

Revéret, L., G. Bailly and P. Badin (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. International Conference on Speech and Language Processing, Beijing - China: 755-758.

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

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

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.

Vertegaal, R., R. Slagter, G. van der Veer, and A. Nijholt. (2001). Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes. Conference on Human Factors in Computing Systems, Seattle, Washington, United States, ACM Press New York, NY, USA.

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.