Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent

Antoine Picot, Gérard Bailly, Frédéric Elisei, and Stephan Raidt

GIPSA-Lab, Dept. of Speech & Cognition, UMR 5216 CNRS/INPG/UJF/Stendhal
46 av. Félix Viallet, 38031 Grenoble - France
Corresponding author: gerard.bailly@icp.inpg.fr

Abstract. We present here a system for controlling the eye gaze of a virtual embodied conversational agent able to perceive the physical environment in which it interacts. This system is inspired by known components of human visual attention system and reproduces its limitations in terms of visual acuity, sensitivity to movement, limitations of short-memory and object pursuit. The aim of this coupling between animation and visual scene analysis is to provide sense of presence and mutual attention to human interlocutors. After a brief introduction to this research project and a focused state of the art, we detail the components of our system and confront simulation results to eye gaze data collected from viewers observing the same natural scenes.

Keywords: embodied conversational agents, face-to-face interaction, eye gaze, talking face, visual scene analysis.

1 Introduction

We produce around 250.000 saccades per day. The eyes of the authors of this paper cover approximately 7m per second when screening computer screens. Multiple factors influence the intensive activity of our gaze control system: perceptive salience of various elements of our field of view (color, shape, motion, etc), their pertinence according to the purpose of the current scan (searching for a particular object or face, decoding the intentions of a human agent, etc) or the *a priori* knowledge we have on each element of the multimodal scene (familiarity, expectations, etc). The main objective of this work is to determine automatically the successive centers of interest that will likely attract the attention and the gaze of our Embodied Conversational Agent (ECA) observing a dynamic natural scene. We particularly propose a gaze control system that identifies and tracks regions of interest, weights their salience and pertinence regarding to the cognitive task, handles a stack of attention and couples visual analysis with an effective gaze control.

Such a strong coupling between a detailed multimodal scene analysis and motor control is necessary for developing ECA sensitive to changes of their real or virtual environment. The environment includes of course the interlocutors: the objective of this grounding of cognitive states and actions is to give to human partners tangible signs of presence and awareness. These cues have an important impact on information processing during interaction in terms of comprehension, belief and cognitive load.

After a brief state of the art where we will detail two major contributions that have inspired this work, we will describe our own proposal and illustrate its properties with concrete examples. This technical presentation is followed by a comparative evaluation with eye-tracking data collected on human subjects.

2 State of the art

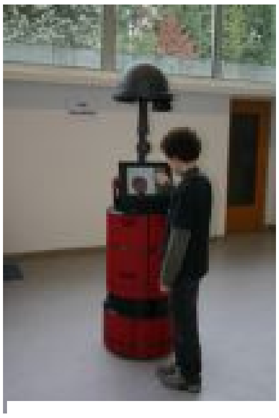

Fig. 1. The Rackham robot interacting with a child (© LAAS Toulouse)

In order to plan their displacements, mobile robots have multiple sensors to build and analyze representations of their surrounding environment. Designing human-aware planning strategies is now a very challenging issue [1]. Most of anthropoid robots or companion robots are sophisticated scene analysis systems (mostly using vision) to analyze human behavior, identify their activities and plan adequate motor responses. Social robots developed at MIT [6] control sensory-motor loops where mutual attention is essential for acquiring and maintaining a common representation space. Maintaining eye contact as well as moving the head and eyes to signal interest or desire to take or leave turn are essential cues for signaling that the loop is effective. Robots developed at Waseda University exhibit such multimodal attention: Robita for example is able to follow a multi-speaker conversation, signal with head and gaze movements that it effectively tracks turns and is thus in a position to take part in the conversation [18]. It is also able to understand and generate multimodal deictic gestures.

Most virtual conversational agents have often poor or none information on the actual environment where the interaction takes place. In absence of a grounded perception, their control model of gaze is often based on statistical regularities such as probability density functions of blinking frequency, amplitudes of ocular saccades or durations of eye fixations. Lee et al. [16] have thus proposed a statistical control model that takes into account the current cognitive activity of the talking agent (notably listening vs. speaking). If the generated saccades are preferred to a fixed or random gaze, the model should benefit from a finer description of the cognitive activities [see 4 and Raidt et al, this volume] as well as an effective coupling with a scene analysis. Proposals made by Courty [10] for virtual scenes or by Gu and Badler [12] for static natural scenes do not address exactly the problem of scrutinizing dynamic natural scenes.

The robot Rackham, developed by LAAS in Toulouse, combines advantages of sense of presence obtained by the performative actions of its body and the communicative actions of our virtual talking face displayed on a screen embedded in the robot (see Fig. 1.). A first coupling of the multimodal scene analysis performed by Rackham with the gaze controller of the talking face has been performed and evaluated [9].

The objective of the present work is to endow this coupling with a more sophisticated gaze controller, capable of reproducing essential characteristics of human visual attention (this paper) and face-to-face interaction (Raidt et al, this volume).

2.1 Behavioral data

The process by which our eyes explore our field of vision consists in a series of saccades and fixations. A fixation here includes an optional smooth pursuit occurring when fixating a moving region of interest. Saccades are rapid movements of the eyes (approx. 25-40ms, 200°/s) that bring the region of interest (ROI) in the central receptive field of the retina (fovea) for further high resolution spectral analysis. Saccades are thus followed by a first fixation (approx. 300ms) followed by corrective fixations [or refixations as amplified by 27 in reading] or a smooth pursuit in case of a slowly moving object of interest. These two main components of gaze trajectory correspond coarsely to two complementary cortical pathways: a dorsal-temporal "where" pathway responsible for localizing multisensorial events in the scene – often termed as the fly detector - and a ventral-parietal "what" pathway responsible for object identification [19]. Jeannerod [14] prefers a more specific differentiation between a pragmatic (perception for action) vs. a semantic (perception for comprehension) analysis of the scene.

Scrutinizing a scene (still image or video) does not only consist in producing saccades towards the most salient object to the next: cognitive demands have a strong impact on the gaze trajectory and object selection [28]. Vatikiotis-Bateson et al [26] have also shown that eye movements during audiovisual speech perception are also influenced both by the comprehension task and environmental listening conditions (signal-to-noise ratio). Attention mechanisms have also a strong impact on scene analysis [see 24 for impressive experiments on inattention blindness].

2.2 Computational models of visual attention and scene analysis

Several computational models of visual attention and scene analysis have been proposed to mimic behavioral data. Numerous models for computing salience maps for still images and videos have been proposed to analyze, encode or summarize visual scenes. We present briefly two models that have attracted our attention because they both have exogenous and endogenous pathways while offering very complementary approaches (cf. Fig. 2.).

Itti et al [13] propose a neurobiological model of visual attention for the control of the movement of the head and eyes of a video-realistic avatar. This model has three main components: (a) a map that associates a degree of intrinsic salience to each pixel of an image by combining several elementary salience maps (movement, orientation, intensity, and color) computed at different scales (using a pyramidal decomposition of the image); (b) a pertinence map that weights this previous map according to the cognitive demand [for an updated proposal, see 20] and (c) an attention map that takes in charge the sequencing of interest points computation by inhibiting zones of interest already scanned (often called inhibition of return mechanism or IOR).

The model of visual attention proposed by Sun [25] is based on a prior hierarchical segmentation of the scene into objects. Elementary processing units are not pixels but segments. The model also performs a syntactic scene analysis organizing segments by salience (from the most salient segment at the largest scale to the smallest) and embedding (from the object to its constituents). Sun also introduces a temporary IOR that restores attention when the appearance of a given segment changes.

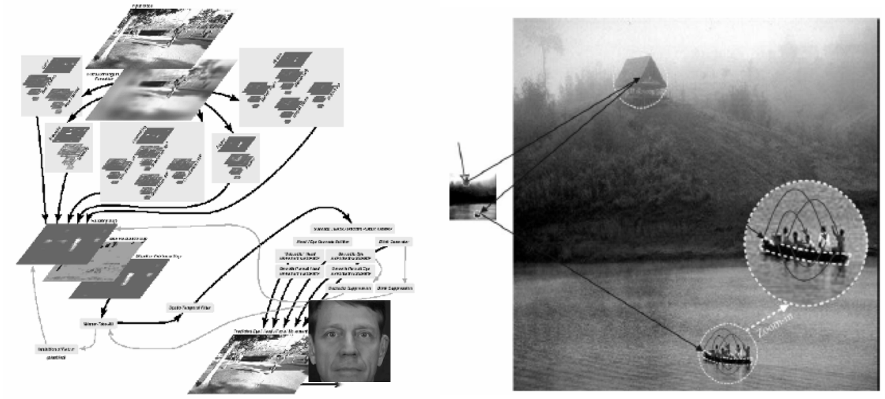

Fig. 2. Models for observing natural scenes. Left: eye saccades of the ECA developed by Itty et al [13] are sequenced by points of interest computed from a video input. Right: Sun [25] uses a multi-scale segmentation to scrutinize an image by successive zoom-ins and -outs.

3 Our model of visual attention

Similarly to Itti et al., the front-end of our model (see Fig. 3.) is a saliency map without prior segmentation. A segmented object is however attached to the most salient interest point in the image by thresholding locally the map. Descriptors of the texture of the segment (notably in contrast with the surrounding region using linear discriminant analysis) are stored in a stack of attention. They are used to track the segment when in motion and also detect segment changes for temporary IOR.

Visual attention is implemented as a stack of attention that temporarily memorizes the position and textural characteristics of segments previously scanned. The stack can function as a FIFO (First-In First-Out) or a LIFO (Last-In First-Out). The most frequent usage is FIFO: it implements a temporary IOR. Each time a new ROI is detected, it is analyzed by the visual system (gaze is directed to it and a minimum fixation interval is planned for object recognition) and pushed on the stack. When the stack is full, the oldest ROI is popped off the stack and discarded. For determining the next most salient object in the scene, we subtract the saliency of all stored items in the stack from the saliency map. A stored item can thus only be scrutinized again if it has been popped off the stack by new incoming items or if its salience (thus its position or appearance) has changed.

The analysis of the current ROI can however be interrupted in order to process an exogenous stimulus that is particularly salient. In this case the stack functions as a

LIFO: the current ROI is pushed on the stack and popped from the stack once the exogenous stimulus has been processed [11].

Our implementation of temporary IOR consists in reactivating a ROI that has changed compared to its stored characteristics or that gains back focus after its removal of the stack due to its limited storage capabilities (the stack has only 4 slots, i.e. possibility of maintaining attention to only 4 ROIs). We also added a smooth pursuit mechanism based on a ROI tracker using Kalman filtering. A module for recognizing and scrutinizing specific objects has also been added to enhance attention towards ROI with high potential interest such as faces.

Note finally that this system includes an effective coupling between saccade generation and visual analysis: a retinal filtering centered on the current position of the fovea cone is applied to the image before computing the saliency map that is thus sensitive to the eye movements. This differs from Itti et al [13] implementation where the center of the retinal filter is always placed at the center of the screen. We do not assume that the camera is monitored by the eye gaze controller as it is usually the case for anthropoid robots [5].

3.1 Saliency map

The saliency map combines the responses of two image processing modules: (a) a “where” module combines orientation (0° and 90°) and movement maps computed at different scales on the raw image and combined by a simple addition; (b) a “what” module combines color and intensity maps computed after the retinal filter. This filter convolves the raw image with a Gaussian filter centered on the current convergence of the eye gaze. The resulting image is blurred according to the distance of the pixel to the center of the field of view. The final saliency map is obtained by summing these two maps normalized by their respective variance (obtained experimentally using a 10mn video). The pixel that is the next target of visual attention is the pixel with maximum salience (Winner Take All or WTA).

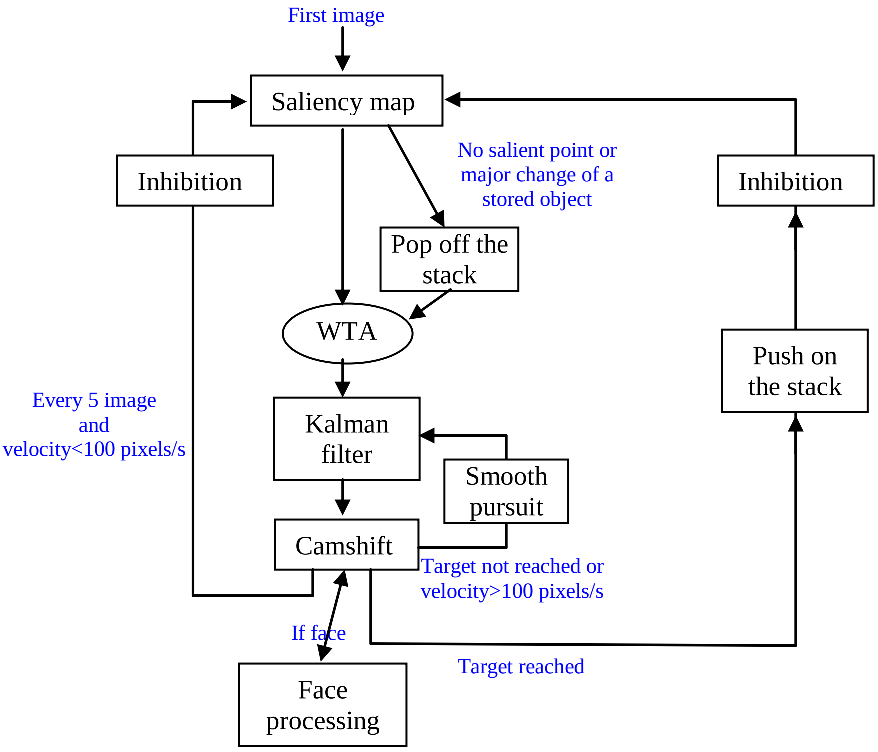

Fig. 3. Synopsis of the proposed system of visual attention. The saliency map is computed using the same scheme as Itti et al [13] except the fact that the retinal filter is centered on the current point of gaze interest.

3.2 Attention stack

Contrary to Itti et al where the interest region is first removed of the attention map and returns slowly to attention using a relaxation process, we adopt an attention stack where the position and characteristics of the current ROI is temporally stored. This scheme is very alike the STM (short-term memory) system proposed by Peters & O'Sullivan [22] except that not all salient objects are stored in the stack and that their locations are effectively stored. The attention drop is just obtained by subtracting the memorized salience of the ROI to the global saliency map. It will thus inhibit the relevant ROI even if its salience is high or will reactivate it if its characteristics changes. The stack has only 4 elements and functions normally as a FIFO: storing a new ROI pops the oldest stored ROI off the stack. An element is thus popped off the stack either if it is too old or because its characteristics has changed (an object passing in the foreground may for example hide the object in the background stored in the ROI or lightening conditions of the ROI may change and enlighten the object for bringing renewed attention to this area). The stack may also function as a LIFO, when the fixation of the current ROI (average 160ms) is interrupted by the coming out of a very salient object in the scene: the current ROI is pushed on the stack, the new ROI is then processed immediately and eventually pursued, and then the memorized ROI is popped off the stack for finishing the WHAT processing.

3.3 Smooth pursuit

In the Itti et al and Sun's proposals, the only module that underlies a possible smooth pursuit is the movement map: the most salient pixel or ROI is expected to coincide with the object in motion. But this is rarely the case, especially when several objects are moving in the scene: this could result in alternating between several points of attention. We have thus added a special module dedicated to smooth pursuit – based on a Kalman filter – that updates the characteristics of the current ROI. The pursuit stops when the estimated speed of the object reaches a minimum threshold. In this mode, the gaze is anchored to the object trajectory and the pursuit has priority over salience computation: the saliency map is only renewed every 5 images in order to be able to process very salient exogenous stimuli (see previous section).

Note also that the control module of the eyes direction uses the estimated speed of the object to anticipate the next position of the object.

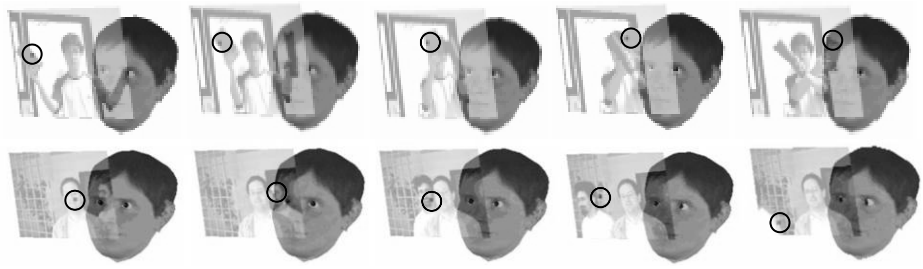

Fig. 4. Our ECA scrutinizing natural scenes. For sake of presentation, the image currently processed is incrusted in a semi-transparent screen placed in front of the ECA. A black circle materializes the current interest point determined by our model of visual attention. The results are given for a few key images of two videos. Top: the subject waves a book in front of the ECA and the smooth pursuit module takes in charge the gaze controller. Bottom: another subject passes in the back of the interlocutor and triggers a saccade to pursue this new and important object of interest.

3.4 Specific objects: face detection

Human vision system is face-aware: the neural activity in a specific zone of the temporal lobe significantly increases when we observe faces (even truncated) *vs.* images without faces or with destructured faces [21]. Such hardwired detectors enable us to focus rapidly on objects of the scene that are semantically very important. The saliency map receives thus a third input: a face detector [we used the built-in OpenCV detector based on 17]. Eyes and mouth detectors are also used [15] to verify the hypothesis and trigger a face-aware gazing of the main elements of the face [26, see also Raidt et al, this volume]

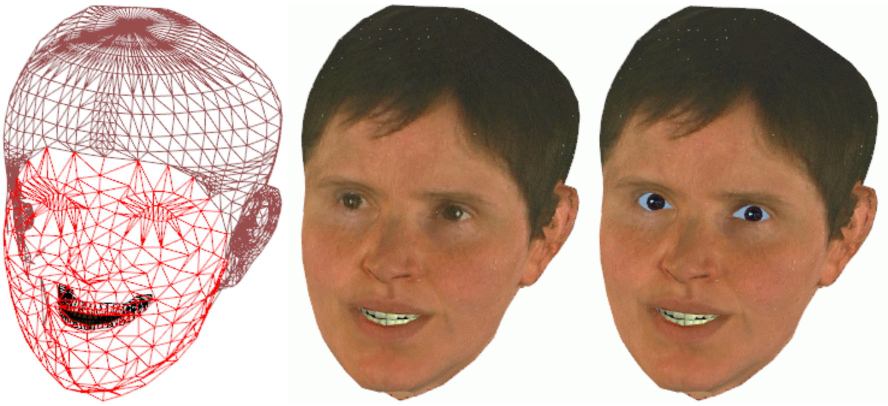

Fig. 5. Our virtual talking face is driven by 12 facial degrees-of-freedom [2]. The eyes and eyelids movements are controlled by 5 degrees-of-freedom that capture the correlations between gaze and eyelids deformations [3].

3.5 Results and evaluation

We present here results obtained on two live videos. The model of visual attention has been coupled with a model of control of the eye gaze of an ECA [3, 7] (see also Fig. 5) that takes care of binocular coordination, saccade generation and micro-saccades during fixations. The trajectories generated by this coupled control (see Fig. 4) have been compared with eye tracking data collected on 5 viewers. Subjects were instructed to view the same videos for further description of each scene to the experimenter. Fig. 4 shows the superposition of the screen coordinates of computed and captured centers of attention.

In Scene 1, one subject first waves a blue book in front of him with his right arm, then a red book with his left arm and then the two books with both arms. This first scene aims at validating our system handling smooth pursuit and the attention stack (since we can only gaze at one object at one time!). Fig. 4 illustrates the results when the subject moves the two books at the same time: the gaze follows the blue book while the red book is ignored and will never be stored in the stack. This inattentional blindness [24] is also present in the occulometric data. The major differences between computed and observed gaze (see Fig. 6) is due to reaction time (our system detects changes more rapidly than human observers) and pursuit of movements (our system pursue the arm gesture although the carried object is outside of the field of view whereas the human viewers exploit causal links between these two movements and switch rapidly back to the subject’s face).

In Scene 2, one subject in the foreground faces the camera while several other subjects walk in the background, stop behind his shoulders and look at the camera. This scene aims at validating our face detector and the stack in LIFO mode (the examination of the face of the main subject is often interrupted by the other subjects passing through). Fig. 4 illustrates the push and pop of the main ROI of the scene to treat an interruption.

The major differences between computed and observed gaze (see Fig. 7) concern the ordinates: the saliency map is too sensitive to saturated colors of the T-shirts that override the salience provided by the face detector.

3.6 Performances

This algorithm has been implemented in C language under Linux Red Hat 9. It uses the OpenCV 0.9.9 Intel library. The tests have been run using a 2002 Pentium 4 desktop at 3.2 GHz. We are close to real time: the processing rate is close to 0.08s per image i.e. 12 images per second.

4 Conclusions and perspectives

We described here a system for scrutinizing natural scenes and its coupling with a controller of the gaze of an ECA. Original components for this model of visual attention have been proposed and implemented: a stack of attention, an integrated face detector and a module for smooth pursuit. This system has been confronted to natural scenes and its prediction has been compared with oculometric data. This confrontation has shown the efficiency of the system in predicting a large part of the observed human behavior. It has also shown limitations and tracks for improvement, notably the importance of top-down processes, cognitive processes and *a priori* knowledge. Mirroring the attention stack and the saliency map, we are planning to add an intention stack [8] and a pertinence map [20] so that the ECA can concentrate its attention to the objects and events directly concerned with its cognitive tasks. Both will determine what to do with salient objects: task-dependent irrelevant ROI should be discarded from the saliency map with a different mechanism than the stack of attention. The pertinence map provides an efficient way to smooth out unimportant ROIs.

A long-term memory component [22] should be also added to feed the intention stack with desired characteristics of the search (location, aspect, etc) for matching objects. The face detector is part of this component: faces and facial expressions are in fact expected to bring more information and comprehension of the scene than other salient ROIs. We have already considered its coupling with a model of mutual attention developed for face-to-face conversation (see Raidt et al, this volume).

Our objective is to develop a real-time implementation of such a sophisticated model of visual attention to quantify its impact on live situated face-to-face interactions where ECA and subjects are involved in collaborative tasks. We have already shown that a pertinent control of multimodal deictic gestures of an ECA has a strong influence on reaction time [23] and plan to use a similar evaluation paradigm for assessing mutual attention.

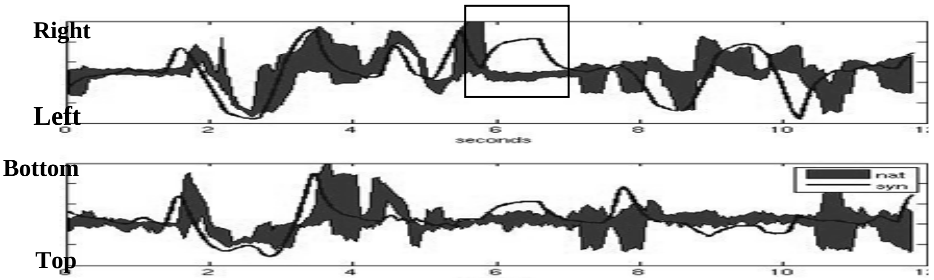

Fig. 6. Comparing the displacement of the gaze focus (top: horizontal; bottom: vertical) predicted by our system with oculometric data (displayed as gauges with means and standard deviations) for scene 1. The major discrepancies are underlined. At t=6s, our system follows the arm movement while viewers go back to the face of the subject (see text for explanation).

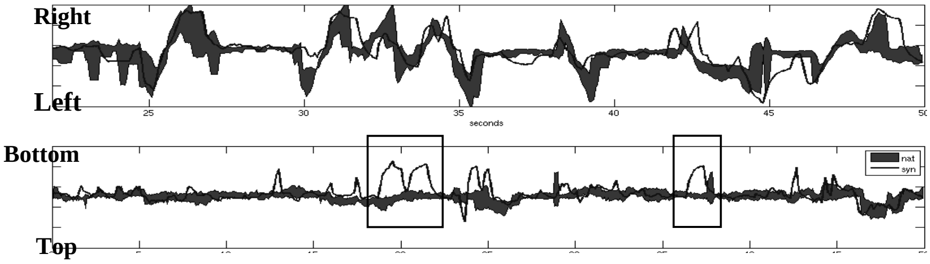

Fig. 7. Same as Fig. 6. but for scene 2. The gaze is often attracted by T-shirts (lower coordinates than faces, see text for explanation).

Acknowledgements. This work has been financed by the project « scrutation de scènes multimodales » supported by ELESA, the project « Deixis Multimodale » supported by GIS PEGASUS, and the project « Présence » of the cluster ISLE supported by Rhones-Alpes. A first implementation of the model proposed by Itti et al has been done by Romain Rossi.

References

1. Alami, R., A. Clodic, V. Montreuil, E.A. Sisbot, and R. Chatila. *Toward human-aware robot task planning*. in *AAAI Spring Symposium "To boldly go where no human-robot team has gone before"*. 2006. Standford.
2. Bailly, G., F. Elisei, P. Badin, and C. Savariaux. *Degrees of freedom of facial movements in face-to-face conversational speech*. in *International Workshop on Multimodal Corpora*. 2006. Genoa - Italy.

3. Bailly, G., F. Elisei, S. Raidt, A. Casari, and A. Picot. *Embodied conversational agents : computing and rendering realistic gaze patterns*. in *Pacific Rim Conference on Multimedia Processing*. 2006. Hangzhou.
4. 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.
5. Breazeal, C., *Designing Sociable Robots*. 2002: The MIT Press.
6. Brooks, R.A., C. Breazeal, M. Marjanovic, B. Scassellati, and M. Williamson, *The Cog Project: Building a Humanoid Robot*, in *Lecture Notes in Artificial Intelligence: Computation for Metaphors, Analogy, and Agents*, C. Nehaniv, Editor. 1999, Springer: New York. p. 52–87.
7. Casari, A., F. Elisei, G. Bailly, and S. Raidt. *Contrôle du regard et des mouvements des paupières d'une tête parlante virtuelle*. in *Workshop sur les Agents Conversationnels Animés*. 2006. Toulouse - France.
8. Chopra-Khullar, S. and N.I. Badler. *Where to look? Automating attending behaviors of virtual human characters*. in *Annual Conference on Autonomous Agents*. 1999. New York.
9. 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. *Rackham: an interactive robot-guide*. in *IEEE International Workshop on Robots and Human Interactive Communications*. 2006. Hatfield, UK.
10. Courty, N., *Animation référencée vision : de la tâche au comportement*, in *IRISA*. 2002, INSA: Rennes. p. 198.
11. Godijn, R. and J. Theeuwes, *The relationship between exognenous and endogenous saccades and attention*, in *The mind's eye: cognitive and applied aspects of eye movement research*, J. Hyönä, R. Radach, and H. Deubel, Editors. 2003, North-Holland: Amsterdam. p. 3-26.
12. Gu, E. and N.I. Badler. *Visual attention and eye gaze during multipartite conversations with distractions*. in *Intelligent Virtual Agent*. 2006. CA.
13. 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.
14. Jeannerod, M., *The cognitive neuroscience of action*. 1997, Oxford, UK: Blackwell. 236.
15. Jiang, X., M. Binkert, B. Achermann, and H. Bunke. *Detection of glasses in facial images*. in *Asian Conference on Computer Vision*. 1998. Hong Kong - China.
16. Lee, S.P., J.B. Badler, and N. Badler, *Eyes alive.* ACM Transaction on Graphics, 2002. **21**(3): p. 637-644.
17. Lienhart, R. and J. Maydt. *An extended set of haar-like features for rapid object detection*. in *IEEE International Conference on Image Processing*. 2002. Rochester - NY.
18. Matsusaka, Y., T. Tojo, and T. Kobayashi, *Conversation robot participating in group conversation.* IEICE Transaction of Information and System, 2003. **E86-D**(1): p. 26-36.

19. Mishkin, M., L.G. Ungerleider, and K.A. Macko, *Object vision and spatial vision: two cortical pathways.* Trends in Neuroscience, 1983. **6**: p. 414-417.
20. Navalpakkam, V. and L. Itti, *Modeling the influence of task on attention.* Vision Research, 2005. **45**(2): p. 205-231.
21. Perrett, D., E. Rolls, and W. Caan, *Visual neurones responsive to faces in the monkey temporal cortex.* Exp Brain Research, 1982. **47**: p. 329-342.
22. Peters, C. and C. O'Sullivan. *Bottom-up visual attention for virtual human animation*. in *Computer Animation and Social Agents*. 2003. Rutgers University, New York.
23. Raidt, S., G. Bailly, and F. Elisei. *Does a virtual talking face generate proper multimodal cues to draw user's attention towards interest points?* in *Language Ressources and Evaluation Conference (LREC)*. 2006. Genova, Italy.
24. Simons, D.J. and C.F. Chabris, *Gorillas in our midst: sustained inattentional blindness for dynamic events.* Perception, 1999. **28**(9): p. 1059-1074.
25. Sun, Y., *Hierarchical object-based visual attention for machine vision*, in *Institute of Perception, Action and Behaviour. School of Informatics*. 2003, University of Edinburgh: Edinburgh. p. 169.
26. Vatikiotis-Bateson, E., I.-M. Eigsti, S. Yano, and K.G. Munhall, *Eye movement of perceivers during audiovisual speech perception.* Perception & Psychophysics, 1998. **60**: p. 926-940.
27. Vergilino-Perez, D., T. Collins, and K. Dore-Mazars, *Decision and metrics of refixations in reading isolated words.* Vision research, 2004. **44**(17): p. 2009-2017.
28. Yarbus, A.L., *Eye movements during perception of complex objects*, in *Eye Movements and Vision*, L.A. Riggs, Editor. 1967, Plenum Press: New York. p. 171-196.