Modeling Dynamic Perceptual Attention in Complex Virtual Environments

Youngjun Kim, Martin van Velsen and Randall W. Hill, Jr.

Institute for Creative Technologies
13274 Fiji Way, Suite 600
Marina del Ray, CA 90292, USA
{yjkim, vvelsen, hill}@ict.usc.edu

Abstract. An important characteristic of a virtual human is the ability to direct its perceptual attention to entities and areas in a virtual environment in a manner that appears believable and serves a functional purpose. In this paper, we describe a perceptual attention model that integrates perceptual attention that mediates top-down and bottom-up attention processes of virtual humans within complex virtual environments.

1 Introduction

In a landscape of ever increasing rendering and animation capabilities there has been an accompanying drive for realistic interaction with intelligent virtual humans. An important characteristic of a virtual human is the ability to direct its perceptual attention to objects and locations in a virtual environment in a manner that appears plausible as an overt behavior and also serves a functional purpose. Not only must virtual humans pay attention to objects related to the tasks they are performing, but they must also be able to cope with sudden events that demand attention. It is often the case that the amount of information in the virtual environment far exceeds the processing abilities of the virtual human. In fact, only a small fraction of sensory information can be fully processed and assimilated into the cognitive model. This situation has been exacerbated as the fidelity of the graphical information in virtual scenes has increased—there is a growing lag in the ability of virtual humans to cope with the amount of environmental data presented to them. Fully simulated virtual humans experience similar cognitive loads as humanss, especially in complex, shared virtual environments. One might even expect our digital counterparts to make the same mistakes as we do and might reject incorrect behavior however logical it is considering the circumstances. A large amount of interaction is negotiated through a common experience and understanding of our physical environment. If artificial intelligence progresses to the point where an agent can make human-like decisions, it will still need to make these decisions based on what it perceives to be its environment. Greater power in rendering capabilities expressed in polygons per second can enhance our own visual experience but might not be beneficial to an agent and can even downgrade its performance. A solution to this dilemma can be found in

the human realm. Spatial cognition and especially spatial attention has allowed humans to make sense of the sensory storm that greets us when wake up in the morning.

Computational models of perceptual attention generally fall into one of two camps: top-down and bottom-up. Biologically inspired computational models [4,9] typically focus on the bottom-up aspects of attention, while most virtual humans [1,2,3,5,13,15] implement a top-down form of attention. Bottom-up attention models only consider the image information (e.g, color, intensity, orientation, and motion) without taking into consideration saliency based on tasks or goals. As a result, the outcome of a purely bottom-up model will not consistently match the behavior of real humans in certain situations. Models like Itti's [9] can predict the bottom-up salience of features in an image at any point in time, but such a model is not sufficient to predict where to actually look. Humans are generally task-oriented, and it is safe to say that a great deal of one's time is spent looking at objects related to the current task.

Modeling perceptual attention as a purely top-down process, however, is also not sufficient for implementing a virtual human. A purely top-down model does not take into account the fact that virtual humans need to react to perceptual stimuli vying for attention. For instance, Chopra-Khullar and Badler [2] built one of the most extensive models to date, a psychologically motivated framework for generating the visual attending behaviors of an animated human figure. Their implementation generates believable animation behaviors for a virtual human performing a fairly scripted set of tasks, but it is not clear how the model would fare in a much more dynamic environment where the need to react to events in the world is much higher than the virtual world they describe. Top-down systems typically handle reaction to perceptual stimuli in an ad hoc manner by encoding special rules to catch certain conditions in the environment. The problem with this approach is that it does not provide a principled way of integrating the ever-present bottom-up perceptual stimuli with top-down control of attention.

In this paper, we present a computational model of perceptual attention for virtual humans. This model extends a prior model of perceptual resolution [6,7] based on psychological theories of human perception. This model allows virtual humans to dynamically interact with objects and other individuals, balancing the demands of goal-directed behavior with those of attending to novel stimuli. This model has been implemented and tested with the MRE Project [8]. Based on the findings with spatial cognition in the MRE environment a self-contained software representation was designed termed ASCE (Agent Spatial Cognition Environment) that serves as a rapidly configurable sandbox for experimentation and testing with models of perceptual attention and spatial cognition.

2 Modeling Perception in Virtual Humans

Our virtual humans are implemented in the immersive environment called the Mission Rehearsal Exercise (MRE) [8] and in the Agent Special Cognition Environment (ASCE). The virtual humans' behavior in MRE is not scripted; rather, it is driven by a set of general, domain-independent capabilities. The virtual humans perceive events

in the scenario, by interacting with the simulator, reason about the tasks they are performing, and they control the bodies and faces of the PeopleShop™ animated bodies to which they have been assigned.

ASCE allows virtual humans to dynamically interact with objects and other individuals, balancing the demands of goal-directed behavior with those of attending to novel stimuli. Using ASCE we can extend the semi 3-dimensional representation into a full spatial model by taking the inverse of our entity model that represents the available space for navigation.

We have developed a model of perceptual resolution based on psychological theories of human perception [6,7] for virtual humans in MRE and ASCE. Hill's model predicts the level of details at which an agent will perceive objects and their properties in the virtual world. He applied his model to synthetic helicopter pilots in simulated military exercise. We extended the model to simulate many of the limitations of human perception, both visual and auditory.

2.1 Visual Perception

As a human has a visual field that extends to around 95 degrees from the center, we limited the virtual human's visual field of view to 190 horizontal degrees and 90 vertical degrees so that the virtual human only gets updates that he is currently sensing through the field of view (FOV). When the virtual human senses the objects in the FOV, it first processes how salient each object is in the respect of size, distance, and color. We consider the computational model [11] to compute the visual salience of each object that is measured by observing individual visual attributes (e.g., size, shape, and color). After computing the visual saliencies of the perceived objects, we applied a sigmoid function as a utility function that reduces the degree of salience of an object in the respect of angle disparities between the virtual human and the object.

2.2 Auditory Perception

Human behavior is very often influenced by auditory inputs that appear to have automatic access to the eye control system via the lower levels. To model auditory perception, we estimate the sound pressure levels of objects in the environment and compute their individual and cumulative effects on each listener based on the distances and directions of the sources. This enables the virtual humans to perceive auditory events involving objects not in the visual field of view. For example, when a virtual human hears a vehicle is approaching from behind, he can choose to look over his shoulder to see who is coming. Another effect of modeling aural perception is that some sound events can mask others. A helicopter flying overhead can make it impossible to hear someone speaking in normal tones a few feet away. The noise could then prompt the virtual human to shout and could also prompt the addressee to cup his ear to indicate that he cannot hear.

Given a set of visually or aurally perceived objects, the agent's perceptual model updates the attributes of objects that fall in the limited sensory range. At any point in time, the virtual human must recognize which object is the most salient among those

objects and draw his focus of attention on the object. The next section describes our approach to computing the salience of the objects in the field of view and the subsequent behaviors associated with shifting the agent's gaze.

3 Computational Model of Perceptual Attention

To compute object salience and to control gaze behaviors, we have developed a model called Dynamic Perceptual Attention (DPA). Internally, DPA combines entities selected by bottom-up and top-down perceptual processes with a decision-theoretic perspective and then selects the most salient entity. Remember that this entity can be an area in the case of olfactory stimuli. Externally, DPA controls an embodied agent's gaze not only to exhibit its current focus of attention but also to update beliefs (e.g., position) of the selected object. That is, the embodied agent dynamically decides where to look, which entity to look for, and how long to attend to the entity.

3.1 Decision Theoretic Control

One of the consequences of modeling perception with limited sensory inputs is that it creates uncertainty on each perceived object. For instance, if an object that is being tracked moves out of an agent's field of view, the perceptual attention model increases the uncertainty level of the target information of the object that a virtual human tries to observe.

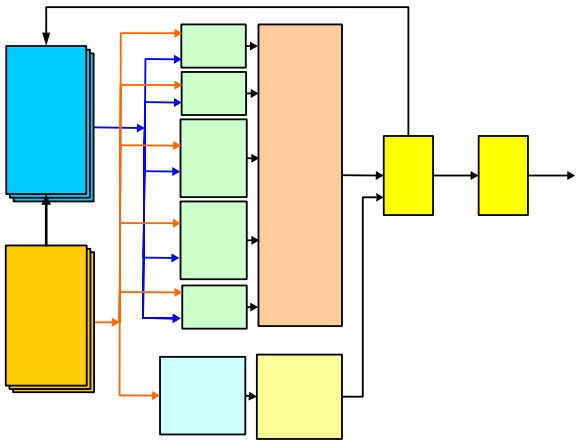

Fig. 1. The Information flow of the DPA module

The information flow of the DPA module is shown in figure 1. Top-down and bottom-up processes give provide information to the DPA module in the form of tuples that are composed of the following components composed as follows:

$$tuple_i = \langle objP_i, objC_i, objDGI_i, objCGI_i, k_i \rangle$$

The priority attribute, objP, is used to indicate the absolute importance of an object, whereas the concern attribute, objC, is used to indicate a conflict between the desired goal information (objDGI) and the current certainty of information (objCGI). For instance, even if a person is given a high priority task, he may not be concerned about monitoring objects associated with the task if the task is going well, resulting in less frequent observations. If the task goes differently what is predicted, he will increase his concern for the task, resulting in more frequent observations.

By considering both attributes (i.e., priority and concern), our virtual humans compute the benefits of attending to objects. Information certainty is one of factors that help the virtual human decide which object it has to focus on. To deal with certainties of the perceived objects, we have chosen to take a decision theoretic approach to computing the perceptual costs and benefits of shifting the focus of perceptual attention of the perceived objects. In the next two sections, we will describe how to compute the perceptual costs and benefits of shifting the focus of perceptual attention. The expected cost is computed by calculating the perceptual cost of shifting the gaze to the selected object. The expected benefit is computed by considering the value of having acquiring accurate information about the selected object. Once a decision has been made, DPA shifts the virtual human's gaze to focus his perceptual attention on the object that has the highest reward.

3.2 Computing the Benefit

To compute the benefit of focusing perceptual attention on an object requires the estimated values of object-based information certainty. We consider object-based information certainty as a key factor in computing the benefit of shifting the focus of attention to the object. The term, *object-based information certainty*, is used here to describe the level of information certainty of an object rendered in the agent's mental image of a virtual world. Humans determine the desired goal information certainty of perceived objects (*objDGI*) based on their subjective preferences or prediction and then make efforts to maintain the current certainty of information (*objCGI)* within a certain specific range of objDGIs, that is defined as the information certainty tolerance boundary (*ICTB*) in our model).

Information certainty is dynamic both in space and time and requires stochastic functions of time and space to describe its dynamics. If the current certainty of information (*objCGI*) is out of *ICTB*, we activate one of two kinds of NEEDs: the NEED for observation or the NEED for inhibition. The NEED for observation is activated if objCGI goes below $ICTB_{lower}$. The NEED of inhibition is activated as objCGI goes over $ICTB_{upper}$. According to Klein's account of the behavior of *inhibition* and *observation* comes from the concept, the *inhibition of return* [12], too

much information can be a bad thing. , which is the process by which the currently attended location or information is prevented from being attended to again and is a crucial element of attentional deployment of humans. By modeling the inhibition of return, perceptual attention will not permanently focus on the most active salient information but will increase the chances of diverting perceptual attention to less salient information.

The orthogonal process model between information certainty and the NEEDs of observation and inhibition is shown in figure 2.

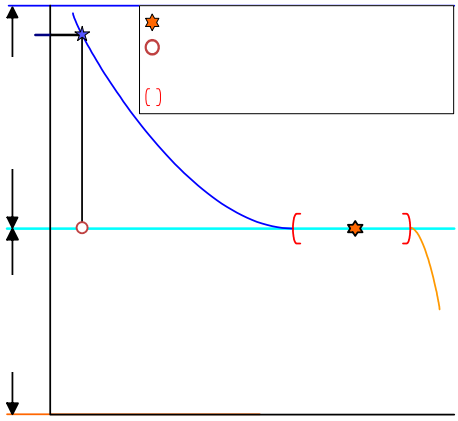

Fig. 2. The interrelation of Information Certainty and Need

The desired goal information certainty (objDGI) is determined by the priority attribute (objP). The information certainty tolerance boundary is set by the concern attribute (objC). The higher the concern attribute is, the narrower the length of the boundary is. The current goal information certainty of the target object (objCGI) is set by top-down and bottom-up processes. If a virtual human cannot retrieve any information certainty of the target from top-down and bottom-up processes, it sets objCGI as 0. After the values for objCGI and information certainty tolerance boundary are set, the virtual human computes the NEED for observation or for inhibition on each tuple as follows:

$$\begin{cases} -1.0 \times objP_i \times \exp^{\alpha} \\ objP_i \times \exp^{\beta} \end{cases}$$

The NEED $tuple_i$ is used as a force that produces a benefit of diverting perceptual attention into $tuple_i$. The benefit is computed as follows:

$$BENEFIT(tuple_i) = \frac{NEED(tuple_i)^2}{2}.$$

Once $BENEFIT(tuple_i)$ is computed, it will used with $COST(tuple_i)$ as the factor to compute $REWARD(tuple_i)$ *with* $COST(tuple_i)$.

3.3 Computing the Cost

Even if the benefit of drawing attention to one object is higher than the benefits of attending to others, the virtual human should not automatically select that object as the best one since the cost of shifting the focus of attention must also be considered.

To compute the cost of shifting perceptual attention from one object to another, we consider two sets of factors: physical and social factors. Physical factors include the degrees of head and eye movements and distance efficiency. Social factors indicate the relative costs of perceptual gaze shifts in social interaction. For instance, it may be rude to look away when someone is speaking (high cost of shift), yet it may be very important to attend to an unexpected or potentially dangerous event (high cost not to shift).

3.4 Shifting Perceptual Attention

With the benefit and two sets of cost factors of each tuple, we compute $REWARD(tuple_i)$ as follows:

$$REWARD(tuple_i) = BENEFIT(tuple_i) - COST(tuple_i).$$

After calculating REWARD(tuple) of all tuples, the virtual human selects a tuple that has the highest REWARD. If the selected tuple is holding the current focus of perceptual attention, the virtual human will keep focus on it. If not, it will divert its perceptual attention into the tuple having the highest REWARD.

The duration of a gaze at an object affects the information certainty level. While a virtual human gazes at an object (i.e., overt monitoring), the objCGI increases. Likewise, while the object is monitored only in the virtual human's memory and projection (i.e., covert monitoring), objCGI decreases. Covert monitoring will cause the certainty of information to decay over time.

4 Perceptual Attention within the MRE Scenario

We implemented dynamic perceptual attention with virtual humans in the immersive environment called the Mission Rehearsal Exercise (MRE) [8].

In MRE, there are three embodied conversational virtual humans – the sergeant (SGT), the mother, and the medic – and a human participant (lieutenant) in an accident site where an Army vehicle has crashed into a civilian car, injuring a boy.

The participant then takes on the task of directing the troops to rescue the boy by interacting with virtual humans. While the rescue task is proceeding, the mom perceives that the troops are moving out of the accident site. In despair, she stands up and cries out for help. Unfortunately, our virtual humans – the sergeant and the medic – are not aware of the mom's outcry, but the human participant is. The system may handle this bottom-up form of attention capture in an ad hoc manner by encoding special rules to catch certain condition in the environment. However, this approach does not provide a principled way of integrating the ever-present bottom-up perceptual stimuli with top-down control of attention.

Traum and Rickel [15] presented an attention layer in a state-of-art model of multi-party dialogue in MRE, which is organized as a set of dialogue management layers. Their attention model is not fully implemented and only (visual) give attention is currently fully modeled. Our aim in this section is to extend their attention layer by addressing the issue of dynamic perceptual attention (DPA). We controlled the sergeant's gaze movements with DPA.

Let's assume that the mom cries out for help while the sergeant is interacting with the medic to talk about the boy's health status. The sergeant's auditory perception computes how loud the mom's outcry is. If the outcry is audible, auditory perception assigns the priority in proportion to the loudness of the sound. Since the sergeant has never expected the auditory input, he will get the certain level of NEED of the auditory input. The auditory perception generates a tuple for the mom's outcry as follows:

$$\neg(mom\_outcry) = \langle objP=0.7, objC=0.0, objDGI=0.8, objCGI=0.0 \rangle$$

The task reasoner gives a tuple for the task of interacting with the medic as follows:

(get_info_boy_health_status)=<objP=0.9,objC=0.0,objDGI=0.7,objCGI=0.5>

The figures for this situation are shown in the figure 3.

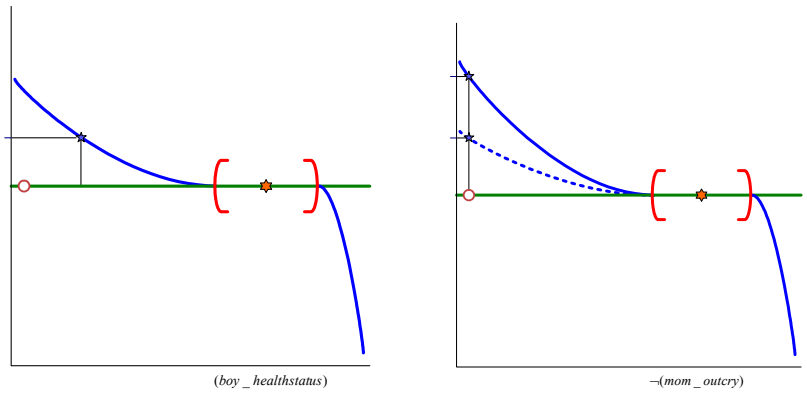

Fig. 3. Information Certainties and NEEDs of (boy_health_status) and $\neg(mom\_outcry)$

When DPA computes the REWARD of each tuple, the bottom-up auditory stimuli, the mom's outcry, gets the highest REWARD and then the sergeant diverts his attention from the medic to the auditory stimuli. This shift of attention will make the

sergeant to recognize that the mom is crying out for help and update the tuple for the auditory stimuli as follows:

$\neg(mom\_outcry)$ =<objP=0.7,objC=0.0,objDGI=0.1,objCGI=0.0>

The recognition of the mom's outcry for help makes the sergeant to interact with the mom so that the sergeant gets what makes her upset. As the result of this recognition, a new tuple is given as follows:

(get_information_from_mom)=<objP=0.7,objC=0.0,objDGI=0.7,objCGI=0.0>

The figures for this situation are shown in the figure 4.

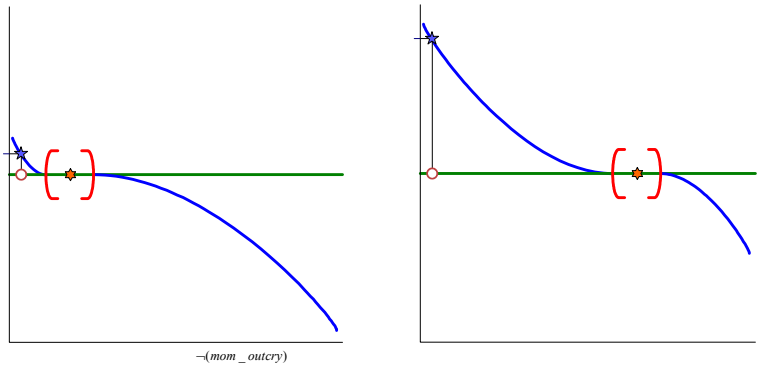

Fig. 4. Information Certainties and NEEDs of $\neg(mom\_outcry)$ and (get_info_from_mom)

The tuple for (get_information_from_mom) now gets the sergeant's focus of attention. While getting information from mom, the sergeant gets certain level of information certainty on (get_info_from_mom).

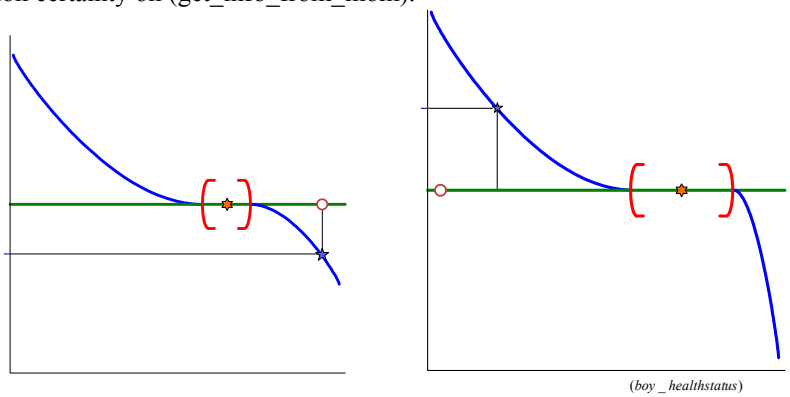

Fig. 5. Information Certainties and NEEDs of $\neg(mom\_outcry)$ and (get_info_from_mom)

If the sergeant understands what she wants, he should think of what he should do to soothe the mom. While thinking, the sergeant updates the tuple, (get_info_from_mom) and gets the tuple, (boy_health_status), as the next focus of attention. The figures for this situation are shown in the figure 5. The sergeant will

HIGH

divert his attention to the medic to update the current state of the medic. As the result of thinking, the sergeant decides to say something to the mom such as “it’s ok, we are staying right now with you. See, we’ve got the medevac coming right now.” This speech event will make the sergeant to divert the sergeant’s focus of attention to the mom and then the sergeant speaks the utterance as mentioned above. After getting the mom’s proper reaction, the sergeant then shifts his focus of attention back to the medic.

This example illustrates the importance of gaze in acquiring perceptual information and monitoring task performance while embedded in the social context of conversation. Our aim is to have the sergeant’s behavior seem appropriate within this context, both in terms of behaving human-like and using perceptual gaze to mediate between costs and benefits of information updating actions.

5 Modeling a Virtual Environment with ASCE

Any artificial environment where synthetic human-like characters represent digital life should also contain an artificial form of the complete range of sensory stimuli.

Within ASCE, where the environment is represented in a discrete fashion, all sensory sensations from the agent's perspective are experienced through interacting with volumetric sensory entities with varying complexity. In order to facilitate spatial perception, we developed a unified sensory oriented environment representation. ASCE’s runtime simulation component analyzes the world geometry and extracts a volumetric representation. In essence, every object is fitted with a bounding volume with associated salience information, creating a stimulus entity. Doing so transforms the geometric model into an existence model, rich with sensory data. This new model can also be seen as a 3-dimensional representation of a saliency map [9]. Only cubic shapes are used for bounding volumes, much like the approached used by Noser et al [10] and Zhang et al [17]. Human perception spans the full spectrum of sensory experiences, from tactile sensations to olfactory stimulations. A volumetric approach provides a computational model that is faster to analyze and maintain than working with the original geometry [17]. Another rationale for this approach is that humans, unlike most robotic and agent implementations, do not consider all details of an entity relevant for interaction. When entering a room filled with objects with the purpose of sitting down at a desk for instance, it is not important what the logo in a trashcan is, it is only important to know its general dimensions and relative placement to avoid colliding with it. ASCE can represent and process any kind of stimuli provided that they are represented as a volume with cubic dimensions [14], out of the entire range of stimuli we will focus here on two of them to show how they are represented and used. One of the improvements over previous approaches [16,7] of data models for artificial sensors is that the inverse of our entity model represents the available space for navigation, which we termed the opportunity space. This allows for higher navigation fidelity than implementations within MRE. A further enhancement is obtained by using spatial attention to only refine those entities that were specifically paid attention to. Doing so removes the limitation of having to use a maximum observation radius [10] in which to calculate volumetric information.

6 Discussion and Future Work

One of the distinctions between the work described in this paper and other models of perceptual attention is the purpose of the model. In many of the systems we reviewed, the purpose of perceptual attention was to make the virtual human behave as though it was attending to the surroundings and tasks in a natural way. In contrast, our goal is also to develop virtual humans that can perform tasks, react to contingencies, interact with other agents, both virtual and human, plan, and make decisions about what to do next or at some future time [6,7]. To accomplish this, we have found that perceptual attention is a critically important mechanism for restricting the sensory information being processed by the perception module and controlling virtual humans to exhibit goal-directed and reactive behaviors. While the model of perceptual attention presented in this paper handles many aspects of behavior generation (e.g., gaze movement), there is another factor in the broader scope of attention: social attention. In a social situation, perceptual attention may interact with social attention since social factors may also change the relative costs of perceptual attention shifts. For instance, it may be very rude to look away when someone is speaking (high cost of shift), yet it may be very important to attend to an unexpected or potentially dangerous event (high benefit to shift). With high utility on either end, the choice may be difficult and moreover potentially very costly either way. By integrating an efficient social attention model with this model, we believe that the model will provide a large potential for generating more socialized behaviors.

7 Acknowledgment

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Cassell, J., H. Vilhjalmsson: Fully Conversational Avatars: Making Communicative Behaviors" Autonomous Agents and Multi-Agent Systems, Vol. 2. Kluwer Academic Publishers (1999) 45-64
2. Chopra-Khullar, S., Badler, N.: Where to Look? Automating Attending Behaviors of Visual Human Characters. Autonomous Agents and Multi-Agent Systems. (2001)
3. Conde, T., Thalmann, D.: An Artificial Life Environment for Autonomous Virtual Agents with multi-sensorial and multi-perceptive features. Computer Animation and Virtual Worlds , Volume 15, Issue 3-4, John Wiley (2004)
4. Courty, N., Marchand, E., Arnaldi, B.: A New Application for Saliency Maps: Synthetic Vision of Autonomous Actors. IEEE Int. Conf. on Image Processing, ICIP'03, Barcelona, Spain, Sep. (2003)
5. Gillies, M., Neil, D.: Eye Movements and Attention for Behavioural Animation. The Journal of Visualization and Computer Animation. (2002)

6. Hill, R.: Modeling Attention in Virtual Humans. Proceedings of the 8th Conference on Computer Generated Forces and Behavioral Representation, SISO, Orlando, Fla. (1999)
7. Hill, R.: Perceptual Attention in Virtual Humans: Toward Realistic and Believable Gaze Behaviors. Proceedings of the AAAI Fall Symposium on Simulating Human Agents, pp.46-52, AAAI Press, Menlo Park, Calif., (2000)
8. Hill, R., Gratch, J., Marsella, S., Rickel, J., Swartout, W., Traum, D.: Virtual Humans in the Mission Rehearsal Exercise System. Künstliche Intelligenz (KI Journal). Special issue on Embodied Conversational Agents (2003)
9. Itti, L., Koch, C.: Computational Modeling of Visual Attention. Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, Mar (2001)
10. Noser, H. and Thalmann, D.: Synthetic vision and audition for digital actors, in Proceedings of Eurographics 1995, 1995
11. Nothegger, C., Winter, S., Raubal, M.: Selection of Salient Features for Route Directions. Spatial Cognition and Computation 4(2): 113-136 (2004)
12. Klein, R.: Inhibition of return. Trends in Cognitive Sciences, 4, 138–147. (2000)
13. Rickel, J., Johnson, L.: Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control. Applied Artificial Intelligence (1999)
14. Samet H.: The Quadtree and Related Hierarchical Data Structures, ACM Computing Surveys (CSUR), v.16 n.2, p.187-260, June (1984)
15. Traum, D., Rickel, J.: Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. AAMAS'02, July 15-19, Bologna, Italy (2002)
16. Young, R. M., Riedl, M.: Towards an Architecture for Intelligent Control of Narrative in Interactive Virtual Worlds, In the Proceedings of the International Conference on Intelligent User Interfaces, January, 2003
17. Zhang, H., Wyvill, B.: Behavioural Simulation in Voxel Space. Computer Animation '97, June 04 - 07, Geneva, SWITZERLAND (1997)