

Perceptual Attention in Virtual Humans: Towards Realistic and Believable Gaze Behaviors

Randall W. Hill, Jr.
University of Southern California
Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292-6695
310-448-8783
hill@isi.edu

1. Introduction

Virtual humans, like real humans, need to perceive their environment. They need to be aware of situations in their world so that they 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. Humans develop a knack for directing their gaze toward the right objects in the environment at the right time—virtual humans need the same kind of ability. But for an increasing number of applications, a virtual human's gaze behaviors must go beyond serving situational perceptual needs—they must also be realistic-looking to human observers. This is particularly true in social situations, where the direction of one's gaze plays a significant role in the communication (Cassell and Vilhjalmsson, 1999; Argyle and Cook, 1976).

Is it possible to generate believable gaze working from a principle-based model? This paper explores the question of how a virtual human's gaze can be driven by an intrinsic model of perceptual attention and cognition while at the same time exhibiting believable behaviors. There is a tension between what a model of perceptual attention can generate and whether it is believable when observed by a human. At one extreme, animated characters can portray a believable human character and yet have no autonomy or cognition at all. Such a character is purely a graphics animation that has been given emotional expression, gesture, and behavior by a human animator, who decides every nuance of the performance. The problem with this approach, however, is that characters cannot dynamically interact with the environment or with other agents. At the other extreme, a virtual human could potentially be developed with a very realistic model of perceptual attention, but fail to create a believable human character, since gaze behaviors are purely functional—they may be socially or emotionally moderated.

2. Motivation

Why should we bother attempting to build models that are at once realistic and believable? Part of the answer to this question was addressed by a National Research Council (NRC) study headed by Pew and Mavor (1998) on modeling human and organizational behavior. The focus of the study was on how to make these models more realistic and robust for a variety of uses—training, mission rehearsal, and military tactics evaluation, to name a few. Realism is important for generating outcomes in these simulations since they are usually crafted to represent real world situations. Furthermore, many simulations today are interactive, so the models of humans and organizations must not only be realistic but also robust—they have to be capable of handling many different situations without breaking down due to brittleness. Simulation-based agents are already being used for training (Rickel and Johnson, 1999), games (van Lent and Laird, 1999),

animated characters (Chopra and Badler, 1999; Terzopoulos et al., 1996), entertainment, military simulations for mission rehearsal and tactics evaluation (Hill et al., 2000; Gratch and Hill, 1999) and are potentially useful for testing computational models of psychological theories.

While realistic models are important for generating outcomes, it is also important that the exhibited behaviors are believable.

3. The Psychology of Perceptual Attention

Before entertaining the question of how to generate believable gaze behaviors, let us first consider some of the psychological properties of perceptual attention that underlie the behaviors themselves.

Why is attention needed?

Humans coordinate perception and cognition using a set of mechanisms that enable perceptual attention. Humans need perceptual attention because there is simply too much information in the visual field for the perceptual system to process. Wolfe (1994) points out that there are two ways of dealing with this problem. The first way is to ignore excess information. By its nature, the human eye discards data. Acuity is limited over much of the visual field. The arc of greatest acuity is located in the fovea and is only 1-2 degrees, consequently much of the visual input is simply not sensed. The second way of dealing with too much data is to be selective in processing the information that is sensed—attention does this by focusing the processing capabilities on a specific region. Posner (1980) describes attention as the system for controlling the way information is routed and for controlling processing priorities.

How is perceptual attention oriented?

The metaphor of a spotlight is frequently used to describe how attention is limited to a particular region (or object) and how it can be moved from place to place. The spotlight should not be equated with the eyes, however. Posner (1980) identifies two forms of orienting, covert and overt. Although the fovea provides the greatest degree of acuity, attention is not tied to the fovea and can be devoted to other parts of the visual field. This type of orienting is covert—it moves the spotlight of attention without moving the eyes. The model of attention we use in this paper is covert. Conversely, overt orienting involves moving the eyes, and sometimes the head, to shift attention to a new region or object. Eriksen and Yeh (1985) hypothesize that attention acts more like a zoom lens than a spotlight. According to this metaphor, a zoom lens has a reciprocal relation between the magnification or level of detail and the size of the viewing field. When the zoom lens is set to a low level of magnification, the field of view is greater, albeit with a low level of detail. Conversely, as the power of the lens increases, the amount of detail increases, while the field of view decreases. According to Eriksen and Yeh, an ideal lens produces a constant amount of information across all power settings.

How is attention controlled?

This question differs from the last in that it deals with what prompts the spotlight of attention to move. Attention can be controlled top-down or bottom-up (Egeth and Yantis, 1997; Neisser, 1967; Posner, 1980; Wolfe, 1994). Top-down control is endogenous—it is directed by a decision of the cognitive system and performed in service of a task. This type of control is used to support visual behaviors such as search and tracking. Bottom-up control is exogenous—in this case, attention is captured by an external stimulus. Stimuli that have been shown to capture attention include luminance, color, motion, and abrupt onset (Neisser, 1967; Wolfe, 1994).

Stages of Perceptual Processing

Visual information is processed in preattentive and attentive stages (Neisser, 1967; Julesz and Bergen, 1983; Wolfe, 1994). Preattentive processing is thought to operate in parallel across the entire visual field; it is the stage where segmentation takes place and textures are formed. The products of preattentive processing are the units or objects toward which attention is subsequently directed. Some researchers hypothesize that Gestalt grouping also takes place during the preattentive stage, although these patterns may not be encoded in memory without attention (Moore and Egeth, 1997). Grouping is performed based on proximity, similarity, and motion.

So, what is attention, anyway?

Perceptual attention is not a monolith that arbitrarily decides what to look at and what to process. The converse appears to be true—attention is a set of mechanisms that integrates the perceptual and cognitive systems. The early vision system gives some features in the visual field greater saliency than others, which draws attention. The sudden onset of a stimulus, motion, and sharp contrasts can all capture attention under the right conditions—the human perceptual system is interruptible by design. When a potentially important event occurs in the visual field then attention can be drawn from one region or object to another. All these mechanisms support bottom-up attention, which is important for survival—reactive behaviors can kick in without any deliberate planning. But much of the human activity we wish to model is goal-driven, consequently there are also top-down mechanisms for focusing attention. Thus, attention forms the nexus of perception and cognition. The cognitive system needs what the perceptual system offers—visual objects with features such as location, color, orientation, and motion. Without these percepts, the cognitive system would not know about the external environment. At the same time, the human perceptual system requires the focus and control provided by attention to allocate limited processing resources to a selected region of the visual field. Without a focus, the perceptual system can be overloaded, with unpredictable consequences for the cognitive system. The cognitive system steers attention, focusing the perceptual resources according to goals and tasks.

4. Perceptual Attention in a Virtual Pilot

We implemented a virtual helicopter pilot in Soar, which is both an architecture for constructing intelligent systems and a unified theory of cognition (Newell, 1990). The Soar pilot flies a synthetic helicopter and performs tactical operations with a team of other Soar pilots. Teams of Soar pilots are deployed along with thousands of other entities (i.e., tanks, trucks, individual combatants, airplanes, etc.) in what is known as a joint synthetic battlespace. Perception in the joint synthetic battlespace involves four distinct problems: perception of terrain, perception of messages, perception of cockpit instruments, and the perception of entities. For this paper we focus on the issue of entity perception. Entity perception is driven by the arrival of a stream of entity-state updates. Each update characterizes the momentary state of an entity: it provides information about the identity, location, and velocity of an entity, such as a tank. These updates are filtered through models of the pilot's visual sensors to determine what information is potentially perceptible. Entities that are too far away will be imperceptible. Entities within the perceptible range of the model may still be rendered imperceptible if they are occluded by a terrain feature or an environmental factor such as smoke or dust. The sensor models also determine the resolution of the percept based on factors like distance, dwell time, and visibility. Hence, an entity may initially be recognized only as a vehicle when perceived at a great distance, but it may be identifiable as a specific tank model at a closer range. Given that the state

information of the perceptible entities is directly available to the virtual pilot, many of the standard vision problems can be finessed. For instance, understanding whether a new update refers to a known entity, or whether it represents a newly perceived object is not a problem. Each entity has a unique identifier in the simulation that can be used to associate the entity state information with visual objects in the pilot's memory. Thus, it is simple to resolve new percepts with previously observed entities. While this is probably not a realistic model of how humans resolve visual percepts, it does provide the human functional capability to recognize entities it has seen before.

One of the first mechanisms we implemented in the pilot was a way of orienting attention based on Eriksen and Yeh's (1985) zoom lens metaphor. This mechanism enables the pilot to perceive the environment at different resolutions. At a low resolution the pilot perceives groups of other entities, and at high resolution the pilot perceives the details of individual entities. With an attention mechanism, the pilot exercises greater control over the amount of information that is processed, ameliorating the overload problems experienced in versions of the pilot without attention (i.e., the pilot was attending to everything.) At the same time, the groups perceived at low acuity provide the right level of abstraction for tracking and reasoning about groups of entities. Furthermore, groups play an important role in controlling and shifting the focus of attention.

Group objects are formed by the virtual pilot's perceptual system in one of two ways: automatically or voluntarily. Automatic grouping is initiated by the perceptual system during the preattentive stage of processing, which takes place during the Soar input phase. Groups are formed based on proximity and similarity (e.g., same vehicle type), which are commonly used Gestalt principles for grouping. Each decision cycle, newly sensed visual objects are compared to existing groups—if a visual object is within 500 meters of the center of mass of a group, and its attributes are similar to the group's members, then it is clustered with that group. Otherwise, a new group is formed. New group objects are added to working memory and pre-existing group objects are updated at the end of the input phase.

Groups are dynamic in nature: they form, split, merge, move, and change shape. For this reason, perceptual grouping must also be dynamic. The attributes of group objects are updated every input phase. If members of a group have changed their positions, then the center of mass has to be re-computed along with all the other geometric relationships.

As in humans, the virtual pilot's attention can be controlled in two ways, endogenously and exogenously. To control attention endogenously, which is a top-down, goal-driven form of control, the pilot normally chooses to perceive only groups, which is the lower acuity mode of perception that saves the cost of processing visual objects at high resolution. Groups provide cues such as location, size, and some details about the membership that can be used for searching for specific visual objects. The pilot may be interested in focusing on visual objects with a particular set of features. For example, one of the highest priorities in the pilot's visual search is to identify enemy air defense vehicles—this is driven by a goal of survival. Lower priorities for visual search include enemy tanks and armored vehicles, particularly those that are in firing range. To shift attention to a visual object involves specifying a set of features to the perceptual system that will cause the visual object, if it is present in the visual field, to be processed and added to working memory. One or more of these features may be selected: group membership, the distance between the visual object and the viewer, vehicle class (e.g., helicopter, tank, or truck), vehicle type (e.g., T-80, AH-64, and so on), and force orientation (opposing versus friendly force). Although these do not exactly match the basic visual features found in humans, there is a correspondence. Wolfe (1994) defines a basic feature as something that supports

efficient visual search and effortless segmentation. He identifies ten basic visual features: color, orientation, curvature, vernier offset, size, motion, shape, pictorial depth cues, stereoscopic depth, and gloss. These basic features are preattentively processed and used for visual search. In the same way, we use groups as a way of segmenting a scene, and individual visual objects have features corresponding to depth, size, and shape that can be used for searching and attending.

5. Task and Environment-Oriented Visual Behaviors

While the virtual helicopter pilot demonstrates some of the psychological principles for orienting and controlling perceptual attention, it lacks a number of the key elements of a realistic visual system. For one thing, the pilot's direction of gaze was not being controlled at all. In effect, the pilot had a 360-degree field of view and was exercising a form of covert attention by controlling the amount of detail about objects and groups. To make the model realistic, the virtual human's field of view will have to be restricted to a human capacity level. Once the field of view has been restricted, then we will have to add overt attention to control the direction of gaze.

Overt attention controls where the virtual human looks—but what should it look at? The obvious answer to this is that it should look at objects or regions that fulfill a need for information about the current state of the world. If the task, for instance, is to follow another helicopter, then it will need to visually track the helicopter while it is flying. Depending on how frequently it needs updates about a visual object, it may be sufficient to only monitor the object, allowing it to shift its gaze among different objects.

Chopra and Badler (1999) identified a set of common visual behaviors that will serve as a good starting point for a virtual human: visual search, monitoring, limit monitoring, reaching and grasping, visual tracking, motion in the periphery, and spontaneous looking. There are a couple of interesting things to note about these behaviors. First of all, each behavior serves a functional purpose. We look at these in greater detail in a moment, but this means that there should be a principle related to perceptual attention behind each of the behaviors. Perceptual attention should drive the behavior—the behavior's function relates to either a need for knowledge or to the need to react (survive) in the environment.

The second thing to note about Chopra and Badler's behaviors was that each has a physical manifestation that should look human-like. Since their focus was making the behaviors look believable, they did not emphasize building a functional model (i.e., model-driven perceptual attention). Their model assumes that a cognitive system was generating visual tasks that were being placed into a queue for the gaze motor system. Implemented in JACK, the visual behaviors look realistic and fairly natural. But as we consider how to implement each of these visual behaviors in the context of a model of perceptual attention, we return to the original question: is it possible to build a functional model that looks realistic?

Visual Search

When we want to locate an object in the environment, we search for it. Where we search will be based on where we expect to find the object. If we don't know have any expectations, then a systematic pattern of looking will take place. What drives the search is some goal or task that requires information about the object in question. In the Soar cognitive architecture, it is easy to see where such goals come from. To implement visual search in Soar, productions instigate the search when there is a goal to acquire information about the outside world. Once the object is

found, then the search is terminated and the desired attributes of the object are recorded in working memory.

Expectations about where to find an object can come from a mental model of the situation. Zhang and Hill (2000a,b) have investigated the use of situation templates to understand vehicle formations in a synthetic battlespace. When a template is partially matched, then the unmatched slots in the template are sources of expectation that guide where to look for the missing vehicles.

The gaze behaviors that result from a visual search will have to appear natural, moving at the speed and arc that one would expect in human performance. It may be possible to use some aspect of the Kieras and Meyer's (1995) EPIC system to model the eye.

Monitoring

Some tasks require periodic updates on the state of an object in the world. The sampling rate depends on the level of uncertainty and the criticality of the object's state for the task at hand. For example, if the task is to follow another helicopter while flying, it may be necessary to frequently sample the position and direction of the other helicopter since a collision would have devastating effect. On the other hand, monitoring a slow moving truck from high altitude may not require updates since it cannot move very quickly relative to the helicopter. Kim et al, (2000) investigated how to mentally model the motion of simulated vehicles, particularly when their movement is influenced or constrained by terrain features such as roads, lakes, rivers, and mountains. Besides predicting the motion of the vehicle, the model includes a time-decay confidence level on the prediction. The confidence factor tells how long the virtual human can look away from an object between samples.

As with visual search, the decision to monitor an object comes from a Soar goal or task operator. An interesting problem that arises here is how to interleave the monitoring of two or more objects in the environment. This is where having the ability to mentally model an object not currently in the visual field will aid in shifting among objects. Chong (1998) addressed a similar issue when he developed a model of dual task performance, where tasks are interleaved. A similar approach may be useful in this case. In any case, perceptual attention will have to shift at the appropriate time, resulting in a shift in the direction of gaze in the virtual human. The shifts will have to be made in a natural way, with the same kind of timing and sweep.

Limit Monitoring

As an object reaches a state transition point, it becomes necessary to monitor more frequently. Chopra and Badler (1999) use the example of a crossing light that is about to change color—as the expectation grows that the limit is being reached, visual samples are taken more frequently. As mentioned earlier, this requires an ability to recognize or predict that a state transition is about to occur. Some transitions will be easy to understand and predict, such as when dealing with an object that can be represented as a simple finite state machine or by the physics of the world. But predicting the actions of another virtual human may require more complex forms of agent modeling. For instance, predicting that a car will turn left at a certain time will involve recognizing and integrating features of the world such as roads, intersections, turn signals, speed, the agent's goals (if they are known), and so on.

Reaching and Grasping

This visual behavior is needed for making fine adjustments when moving the hand toward an object. The goal is to grasp the object, and feedback from the visual system enables the virtual human to exercise fine motor control.

Visual Tracking

When a virtual human requires constant updates about a moving object, then visual tracking behavior is required. An example of this is a task where a pilot has to follow another helicopter—the follower has to constantly keep the followed in sight. This type of behavior is initiated by a task, so it involves a top-down form of control over attention. Visual tracking is very similar to monitoring except that update cycle is shorter.

Motion in the Periphery

This is an example of where attention would potentially be captured bottom-up by motion in the periphery. There will be cases where the agent suppresses the urge to shift attention to motion in the periphery since attention is more urgently needed elsewhere.

Spontaneous Looking

This is a form of gaze behavior that may not fall neatly into top-down or bottom-up categories. It is a behavior that humans engage in when they are not purposively searching, monitoring, or tracking. It may be driven by a bottom-up saliency map, or there may be a random top-down control strategy similar to visual search. In any case, it makes the virtual human look more natural.

6. Social Uses of Gaze

Argyle and Cook (1976) describe how gaze provides a powerful form of non-verbal communication. More recently Cassell and Vilhjalmsson (1999) have used gaze as an important communicative behavior in their animated characters. Rickel and Johnson (1999) also employ gaze in their tutoring agent, STEVE, who looks at the student in the eye during conversational interaction, and looks at objects in the environment when performing tasks or monitoring the student. These are examples of how the social aspects of gaze can make a more compelling and believable virtual human. But how does this relate to perceptual attention? Argyle and Cook (1976) suggest that there are a lot of reasons why someone will look at the face and eyes of another to get cues about their mood or disposition during a conversation. The outward manifestations of a person's inner state provide insight to the viewer—if the person with whom one is talking is becoming agitated or angry, there may be visual cues that will give warning. If these cues are perceived and correctly interpreted, then the perceiver has a chance to cool down the situation before it gets ugly. This is one motivation for looking at the face of another person, but social gaze is not that simple—it also involves looking away at the appropriate times. A person's emotional state, culture, social position, personality, and other factors can all moderate gaze behaviors.

7. Conclusions

Our group is currently building models of social intelligence, stress and emotion (see Gratch, 2000; Hill et al., 2000). These models will potentially be useful for generating gaze behaviors that are socially appropriate and moderated by emotional state. As we integrate the cognitive, social, and emotional models with the perceptual attention system, we expect to be able to eventually generate behaviors that are both realistic and believable. The next step will be to generate the task and environment-oriented perceptual behaviors from perceptual attention, and then we can integrate the social forms of gaze. Jeff Rickel and I are currently collaborating on a project where we will incorporate gaze and perceptual attention with virtual human being developed for the Institute for Creative Technologies (ICT) at the University of Southern California. The goal of

this project is to build a mission rehearsal system where virtual humans will serve as team members, tutors, and adversaries, so all these issues are quite real to us.

8. References

- M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.
- W. Bacon and H. Egeth. "Overriding stimulus-driven attentional capture," *Perception and Psychophysics*, 55(5):485-496, 1994.
- J. Cassell and H. Vilhjalmsson. Fully Conversational Avatars: Making Communicative Behaviors Autonomous. *Autonomous Agents and Multi-Agent Systems*, 2: 45-64. Kluwer Academic Publishers, 1999.
- D. Chapman. "Intermediate Vision: Architecture, Implementation, and Use," *Cognitive Science* 16, 491-537, 1992.
- R. Chong, Ronald S. (1998). Modeling dual-task performance improvement: Casting executive process knowledge acquisition as strategy. CSE-TR-378-98, University of Michigan.
- S. Chopra and N. Badler. "Where to Look? Automating Visual Attending Behaviors of Virtual Human Characters," Proceedings of the Third International Conference on Autonomous Agents. May 1-5, 1999. Seattle, WA.
- H. Egeth and S. Yantis. "Visual Attention: Control, Representation, and Time Course," *Annual Review of Psychology*, 1997, 48:269-97
- C. Eriksen and Y. Yeh. "Allocation of Attention in the Visual Field," *Journal of Experimental Psychology, Human Perception and Performance*, 1985, Vol. 11, No. 5, 583-597.
- J. Gratch, "Emile: Marshalling Passion and Emotion for Training," Proceedings of the Fourth International Conference on Autonomous Agents, Barcelona, Spain, 2000.
- J. Gratch and R. Hill, "Continuous Planning and Collaboration for Command and Control in Joint Synthetic Battlespaces," Proceedings of the 8th Conference on Computer Generated Forces and Behavioral Representation, Orlando, FL, May 1999.
- J. Gratch, S. Marsella, R. Hill, and George Stone, "Deriving Priority Intelligence Requirements for Synthetic Command Entities," Proceedings of the 8th Conference on Computer Generated Forces and Behavioral Representation, Orlando, FL, May 1999.
- R. Hill, J. Gratch, P. Rosenbloom. Flexible Group Behavior: Virtual Commanders for Synthetic Battlespaces. Proceedings of the Fourth International Conference on Autonomous Agents, Barcelona, Spain, 2000.
- R. Hill, "Modeling Perceptual Attention in Virtual Humans," Proceedings of the 8th Conference on Computer Generated Forces and Behavioral Representation, Orlando, FL, May 1999.
- B. Julesz and J.R. Bergen. "Textons, The Fundamental Elements in Preattentive Vision and Perception of Textures," *The Bell System Technical Journal*, Volume 62, No. 6, July-August 1983.
- Kieras, David E., and Meyer, David E.. An Overview of the EPIC Architecture for Cognition and Performance with Application to Human-Computer Interaction. EPIC Report No. 5.
- Y. J. Kim, R. Hill, and J. Gratch. "How Long Can An Agent Look Away From A Target?," 9th Conference on Computer Generated Forces and Behavioral Representation scheduled from 16 - 18 May 2000.

- C. Moore, and H. Egeth. "Perception without attention: Evidence of grouping under conditions of inattention," *Journal of Experimental Psychology: Human Perception and Performance*, 1997, Vol. 23, No. 2, 339-352.
- M. Posner. "Orienting of attention," *Quarterly Journal of Experimental Psychology*, (1980) 32, 3-25.
- U. Neisser. *Cognitive Psychology*. NY: Appleton-Century-Crofts, 1967.
- A. Newell. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press, 1990.
- R. W. Pew and A. S. Mavor, editors. *Modeling Human and Organizational Behavior: Application to Military Simulations*. National Academy Press, Wash., D.C, 1998.
- J. Rickel and W. Lewis Johnson. "Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control." *Applied Artificial Intelligence*, 13:343-382, 1999.
- D. Terzopoulos, T. Rabie, R. Grzeszczuk. "Perception and Learning in Artificial Animals," Proceedings of the 5th International Conference on the Synthesis and Simulation of Living Systems, Nara, Japan, May, 1996.
- M. van Lent and J. Laird. "Developing an artificial intelligence engine." Proceedings of the 1999 Game Developers' Conference, San Jose, CA.
- J. Wolfe. "Guided Search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, 1994, 1 (2), 202-238.
- A. Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967.
- W. Zhang and R. Hill. A Template-Based and Pattern-Driven Approach to Situation Awareness and Assessment in Virtual Humans. Proceedings of the Fourth International Conference on Autonomous Agents, Barcelona, Spain, 2000.
- W. Zhang and R. Hill. "Situation Awareness and Assessment: Issues and Computational Approaches," 9th Computer Generated Forces & Behavioral Representation, Orlando, Florida, May 16-18, 2000.