PHOTOREALISTIC ATTENTION-BASED GAZE ANIMATION

Laurent Itti, Nitin Dhavale and Fréderic Pighin

Computer Science and Institute for Creative Technologies, University of Southern California

Fig. 1. Sample gaze animations sequence.

ABSTRACT

We apply a neurobiological model of visual attention and gaze control to the automatic animation of a photorealistic virtual human head. The attention model simulates biological visual processing along the occipito-parietal pathway of the primate brain. The gaze control model is derived from motion capture of human subjects, using high-speed video-based eye and head tracking apparatus. Given an arbitrary video clip, the model predicts visual locations most likely to attract an observer's attention, and simulates the dynamics of eye and head movements towards these locations. Tested on 85 video clips including synthetic stimuli, video games, TV news, sports, and outdoor scenes, the model demonstrates a strong ability at saccading towards and tracking salient targets. The resulting autonomous virtual human animation is of photorealistic quality.

1. INTRODUCTION

This study addresses the challenge of automatically endowing virtual agents with realistic gaze behaviors. While much research in facial animation has focused on speech [1], animating the eyes and rigid motions of the head has received less emphasis. In contrast, animals can precisely estimate another's gaze direction, to an accuracy within 4° for humans [2]. Furthermore, gaze plays a crucial role in conveying intentionality and in interacting with social partners [3], making any inaccuracy in its rendering obvious to even the casual observer. While previous studies have proposed models of *how* human eyes move, the main open challenge addressed here is to derive a procedural technique that also realistically selects *where* the eyes should be pointed to, given unconstrained visual inputs.

Our animation technique relies on research in neuroscience: we developed a neurobiological model for the automatic selection of *visually salient* locations in arbitrary video scenes, augmented with eye and head motion dynamics calibrated against human recordings. The model's output animates a photorealistic 3D human face model, posing more stringent criteria on the evaluation of its realism than an artificial creature or impoverished human face model would. The system is tested against a wide range of video clips, and autonomously predicts the motion of the eyes, head, eyelids, and accompanying deformations of the virtual facial tissue. Applications include interactive games, character animation in production movies, human-computer interaction and others.

The following sections discuss related work on gaze animation, our attention and eye/head movement model, the experimental procedure to gather human eye and head movement metrics that calibrate the model, and finally our photorealistic rendering. Our contributions and focus are on three components: developing a visual attention model that selects behaviorally salient targets in any video input; developing motion models for the eyes, head and eyelids that are calibrated against actual human recordings; and building a photorealistic animatable face-model to support gaze animation.

2. RELATED WORK

Approaches to endow avatars with realistic-looking eye movements primarily fall under three categories. A first approach is to augment virtual agents with random eye movements whose dynamics are matched to those of humans. For example, Lee *et al.* [4] recorded human saccades with and eye-tracker, and derived random probability distributions from the data to augment the behavior of an agent with eye movements. This model does not direct the eyes towards specific visual targets, as gaze direction either follows large head movements or is randomly drawn from the empirical distributions. Thus, this study addresses *how* the eyes move, but not *where*.

A second approach uses machine vision to estimate gaze and pose from humans, then driving an agent in real-time from the estimated parameters (e.g., Heinzmann & Zelinski [5]). By definition, this is only applicable to situations where avatars embody humans, such as video conferencing and online gaming, since it is the human whose gaze is being estimated that defines behavior. This differs from our goal to autonomously animate agents.

A last approach uses machine vision to attempt to locate targets of interest in virtual scenes [6]. For instance, Terzopoulos and Rabie [7] proposed an active vision system for animats. Gaze targets are selected based on color signatures, looking for known objects through color histogram backprojection. While this system is limited to objects with known signatures in uncluttered virtual environments, its architecture is particularly relevant. Our more detailed biological modeling of attention and eye movements allows us to extend this paradigm to a wide repertoire of combined eye/head movements in unconstrained environments, containing arbitrary numbers of known or unknown targets against arbitrary clutter.

3. NEUROBIOLOGICAL MODEL OF ATTENTION

Over the past century of attention research, several key aspects of sensory processing have been defined, such as the anatomical and functional segregation between localizing salient targets ("where" dorsal pathway) and recognizing them ("what" ventral pathway) [8]. Computational models of attention have extensively relied on the idea that a centralized topographic *saliency map*, spatially encoding for salience irrespective of feature dimension, may provide an efficient strategy for guiding attention (see [9] for review).

We developed a computational model that predicts the spatiotemporal deployment of gaze onto any incoming visual scene. The

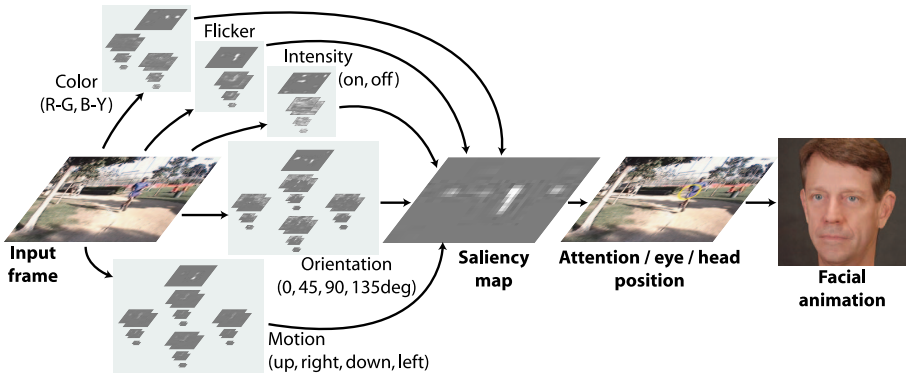

Fig. 2. Overview of the model. Input video is processed by a foveation filter, followed by low-level extraction of multiple visual features (color, motion, etc.) at several spatial scales. All resulting feature maps sum into the saliency map, whose maximum is where the model's covert attention is pointed to. Shifts of covert attention over time drive the overt eye/head movement controller and realistic facial animation.

model is based upon a fairly detailed software replication of the early stages of visual processing in the primate brain, from the retina to higher-level cortical processing areas in the posterior parietal cortex [10]. It is freely available in source-code form.

Incoming video input passes through a foveation filter that blurs each frame in an eccentricity-dependent manner, simulating the highly non-uniform distribution of photoreceptors on the human retina. The foveated input is then processed in parallel by a number of low-level feature maps [10, 11], which detect local spatial discontinuities in color, intensity, four orientations, temporal change, and four motion energies at multiple spatial scales **(Fig. 2)**. Each feature map is endowed with internal dynamics that operate a strong spatial within-feature and within-scale competition for activity [11]: initially possibly very noisy feature maps are reduced to very sparse representations of only those locations which strongly stand out from their surroundings. All feature maps sum into a unique scalar saliency map that guides attention. We refer the reader to previous publications for details [9], while here we focus on generating realistic eye and head animations from the output of this model, i.e., (x,y) coordinates of the most salient location in each video frame.

4. GENERATION OF EYE AND HEAD MOVEMENTS

The so-called *covert* (before eye movements occur) shifts of the spotlight of attention generated by the model provide inputs to the eye/head controller developed in this study. Since covert attention may shift much more rapidly than the eyes, in our model we filter the sequence of covert fixations and elicit an overt saccade if the last 4 covert shifts have been within 10° of each other and at least 7.5° away from current eye fixation.

Behavioral studies of alert behaving monkeys and humans indicate that gaze shifts are accomplished by coordinated motion of eyes and head in the same direction [12]. Sparks [13] provides a comprehensive review on the topic. Of particular interest here, Freedman and Sparks [14] recorded eye/head movements made by unrestrained *Rhesus* monkeys. Remarkably, they found that the relative contributions of head and eyes towards a given gaze shift follow simple laws, at the basis of our model. The total gaze displacement G is the sum of an eye-in-head vector E and a head-in-space vector H, with the relative contributions of eye and head dependent upon the initial angle of the eyes in the orbits. Following Freedman and Sparks [14], for gaze shifts G smaller than a threshold value T, head displacement H is zero. T is given by, with IEP denoting the initial eye position relative to the head and all angular displacements in degrees (IEP is positive if the eyes are initially deviated in the direction of the subsequent movement, negative otherwise):

$$H = 0 \text{ if } -T < G < T \text{ with } T = \left(\frac{-IEP}{2} + 20\right) \times 0.56 \quad (1)$$

For gaze amplitudes outside this zone, total head movement amplitude H and gaze shift are linearly related such that:

$$H = (1-k) \times T + k \times G \text{ with } k = \left(\frac{IEP}{35} + 1.1\right) \times 0.65 \quad (2)$$

This computation is separately carried out for the horizontal and vertical components of the gaze shifts; hence, for a long but nearly horizontal gaze shift, there is some horizontal but no vertical head contribution. Because this model was originally developed for monkey data, we acquired human data in our laboratory to calibrate the eye and head movement dynamics of our model, while we retain the gaze decomposition scheme just discussed.

5. EYE/HEAD MOVEMENT MODEL CALIBRATION

To determine the spatiotemporal characteristics of human eye and head movements, we used a commercial eye-tracker (ISCAN Inc., model RK-464) and a head-tracker developed in-house. With these two devices, we collected descriptive statistics of the motion parameters to be used in the model. Our eye-tracker required that the head be fixed, preventing simultaneous recording of eye and head movements. Hence, we separately recorded eye movements with head fixed, then head movements without eye-tracking. The gaze decomposition model of the previous section was then assumed to apply.

Experiments: Subjects (six for eye tracking, three for head tracking) watched a selection 50 video clips from the 85 used to test the model (next section). For eye tracking, stimuli were presented on a 22" monitor (LaCie Corp; 640×480, 60.27 Hz double-scan, mean screen luminance 30 cd/m^2, room 4 cd/m^2) at a viewing distance of 80 cm ($52.5° \times 40.5°$ usable field-of-view). The eye-tracker was calibrated every five clips. It estimated point of regard (POR) at 240 Hz from comparative tracking of the center of the pupil and the specular reflection of the infrared light source on the cornea. An affine POR-to-stimulus transform was computed, followed by a thin-plate-spline warping to account for any small residual nonlinearity. The head-tracker was custom-built from a bicycle helmet onto which three visual markers were attached. To encourage head movements, subjects were positioned closer to the monitor, yielding a wider usable field of view ($100° \times 77°$). A 30 Hz video camera (Sony Inc.) filmed the helmet while subjects watched video clips with their head unrestrained. On each frame, the three markers were localized using a matched filter, and the 3D head pose was recovered using a quaternion-based method [15].

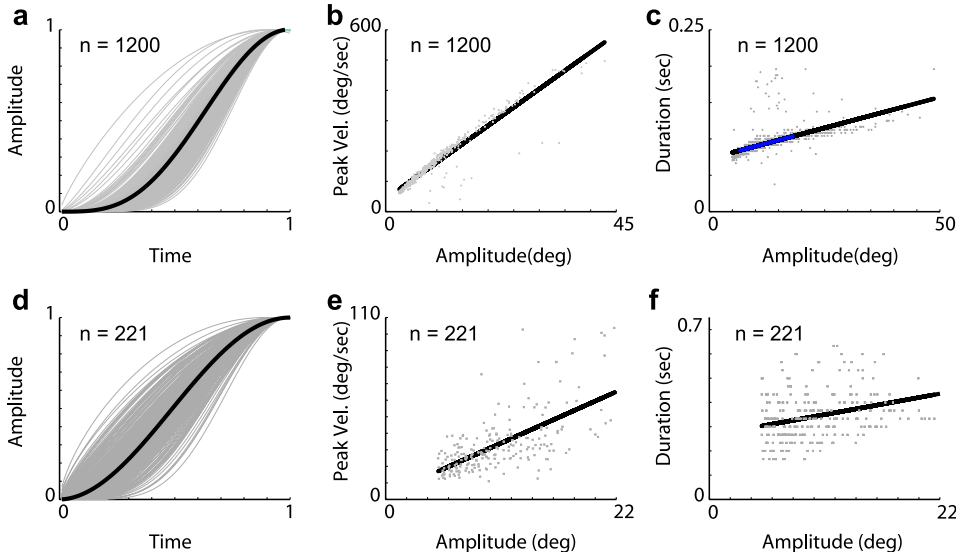

Fig. 3. Eye **(a, b, c)** and head **(d, e, f)** movement data collected in our laboratory while human observers watched video clips on a computer monitor. Least-squares fits to the data yielded models for saccadic velocity profiles **(a, d)**, peak saccade velocity **(b, e)** and saccade duration **(c, f)**. These models were used to calibrate our eye/head movement controller against the human dynamics. Clearly, there are variations from saccade to saccade in our data. Thus, while we our model uses the best-fit curves shown, it is conceivable in the future to also add random variability, to also capture the variance of our observations.

Statistical Data Analysis: Eye and head movement traces were segmented into periods of high-velocity (saccadic) and low-velocity (smooth pursuit or rest) movements. Saccade onset and offset were identified by thresholding smoothed instantaneous velocity. Head movements aborted mid-flight (because a new head movement towards a new target was initiated) were excluded. A total of 1,200 eye and 221 head saccades were thus isolated and analyzed.

Each saccade was normalized to unit angular amplitude and duration (**Fig. 3.a,d**). A model for the normalized angular displacement $\theta(t)$ as a function of normalized time t of the form $\theta(t) = \exp(-kt)\left(\sin\frac{\pi}{2}t\right)^w$ was found to best describe the data, with, for the eye (θ_e) and the head (θ_h):

$$\theta_e(t) = e^{-0.002t}\left(\sin\frac{\pi}{2}t\right)^{3.33} \text{ and } \theta_h(t) = e^{0.027t}\left(\sin\frac{\pi}{2}t\right)^{1.90} \quad (3)$$

We found that both peak eye and head velocities (pV_e and pV_h) varied approximately linearly with the (unnormalized) total saccade amplitudes Θ_e and Θ_h (**Fig. 3.b,e**):

$$pV_e = 12.07\Theta_e + 42.45^\circ/s \quad \text{and} \quad pV_h = 3.45\Theta_h + 2.01^\circ/s \quad (4)$$

Finally, although a linear relationship was less obvious between saccade duration and amplitude, in a first approximation we derived a linear least-squares expression for durations d_e and d_h (**Fig. 3.c,f**):

$$d_e = 0.002\Theta_e + 0.07s \quad \text{and} \quad d_h = 0.010\Theta_h + 0.23s \quad (5)$$

In sum, our data recorded from human subjects allowed us to precisely calibrate the dynamics of the model to human behavior.

Saccadic Suppression: In the primate brain, visual input is inhibited during saccades [16], probably to prevent perception of motion transients as the eyes move. In our model, the best place to implement this is the saliency map; thus, during saccades, the saliency map is entirely inhibited. This has three effects: attention is prevented from shifting; it will take on the order of 200ms for the saliency map to recharge, thus enforcing some intersaccadic latency; and all memory of previous salient visual stimuli will be lost.

Smooth Pursuit: The model of Freedman and Sparks does not include another mode of eye movements found in humans and only a few other animals, by which the eyes can accurately track a slowly moving target using slower eye and head movements [13]. This *smooth pursuit* mode was added to our model, using two mass/spring physics models (one for head and one for eye). When a gaze shift is too small to trigger a saccade, it instead becomes the new anchor point of a zero-length spring linked on its other end to the current eye (or head) position. The head spring is five times weaker than the eye's, ensuring slower head motion.

Eye Blinks: A final addition to our model is an eyeblink behavior, also derived from our data. Except for immediately successive blinks (e.g., double or triple blinks), the probability of a blink occurring during interval $[t..t+1]$ (in s) appeared to decay exponentially with time from the end of the previous blink (**Fig. 4**):

$$P(\text{blink in } [t..t+1] | \text{blink ended at } t_0) = 0.5e^{-0.12(t-t_0)} \quad (6)$$

The blink amplitude distribution was fitted with a Gaussian of mean 190 ms and standard deviation 40 ms. The model's blinks were simulated by sampling from this distribution.

6. ANIMATION AND RENDERING

To convincingly illustrate our eye motion synthesis technique, we map the motions onto an animatable face model **(Fig. 1)**. Contrary to previous automated avatar animation techniques, typically demonstrated with highly impoverished "ball-and-stick" face models or non-human creatures, an important new aspect of our work is the animation of a photorealistic human face model. This sets a much higher standard in evaluating the realism of our results. Our face model is a digital replica of an actor's face. It was constructed and textured from three photographs of a human actor, using a technique similar to Pighin *et al.* [17]. The eyes were modeled as spheres and also textured from the photographs.

The model is controlled through two degrees of freedom for the eyes and two for the head. The head's orientation can be changed according to two Euler angles for elevation and azimuth. The rotation center was estimated from facial motion capture data recorded on the same actor. The eyes have similar controls, and rotate around their geometric center independently of head orientation.

We convert screen coordinates into Euler angles by assuming a field of view of 90°. The head's center of rotation is at the same altitude as the center of the screen. Since we do not know the depths of

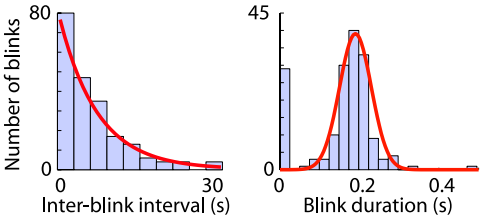

Fig. 4. Histograms of inter-blink intervals and durations.

objects in the videos, we assume that the eyes are focused at infinity and point in the same direction. To make our face behave more naturally, we built a mechanism to animate the eyebrows as a function of the gaze direction, using two blended shapes: The first has the eyebrows level, while the second one has raised eyebrows. During the animation we blend both shapes according to eye orientation.

Because we carefully designed this face model for photorealistic rendering, we produce remarkably convincing animations from our attention model. **Fig. 1** shows a few sample frames.

7. RESULTS

We tested our model on a database of 85 video segments (over one hour playback time in total), including artificial stimuli, natural indoors and outdoors scenes, and video games. Overall, the model attended to locations that made sense to human observers. For example, it well followed the ball, players and overall action in a soccer game, and locked well onto the main character and its enemies in the video games (see companion video). The resulting photorealistic facial animation overall was very convincing.

There are several failure modes of the model in its present form, typically due to its lack of object recognition and scene understanding. Indeed, the model currently attends to the most salient location without knowing its identity. This may yield extended periods during which the system tracks a very salient object that a human would consider irrelevant. For instance, in some of the video game clips, the screen was populated with health indicators that were salient due to bright colors. This indicates that low-level bottom-up saliency, as computed here, is only a factor in the selection of saccade targets in humans. Further work is ongoing in our laboratory to evaluate the behavioral relevance of candidate saccade targets before an eye movement is executed, based on partial identification of the target and evaluation of its relevance to the task at hand.

Nevertheless, our results show that our approach is applicable to an unconstrained variety of stimuli. This contrasts with many computer vision approaches, typically designed for specific targets in constrained environments. Indeed, no parameter tuning nor any prior knowledge of the form or contents of the video clips was used in our simulations, and the exact same model processed all stimuli.

8. DISCUSSION AND CONCLUDING REMARKS

We believe that our model's performance was achieved through our modeling of the neurobiology of attention and eye/head movements, instead of developing a dedicated system for specific environments and targets. Further testing showed a very high statistical correlation between model and human eye movements on the same scenes, which we have carried out in the context of an application to saliency-based video compression [18].

There are many obvious limitations to our model. First, in its present form, it is entirely retinotopic (i.e., all visual processing is made in image coordinates) and does not account for the well-documented coordinate transforms in parietal cortex and other brain areas. Thus, information about previously attended targets is lost as saccades occur. Second, the saliency map is built entirely from the outputs of local operators, although their receptive fields may be large. Adding a global bias for particular locations, based on rapid recognition of the "gist" of the scene, may allow the model to more rapidly orient towards relevant parts of the input [19].

This work is truly a multidisciplinary effort to merge research results from computer graphics and neuroscience. We believe that a fertile synergy between the two fields will result in more accurate and realistic models for graphics, but also will provide validations for theories of low-level human behavior. With this work we not only build a way to realistically synthesize gaze motions but also demonstrate visually the plausibility of our attention model.

Supported by NSF, HFSP and DARPA.

9. REFERENCES

[1] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proc. SIGGRAPH*. 1997, pp. 353–360, ACM Press.

[2] C. Gale and A. F. Monk, "Where am I looking? the accuracy of video-mediated gaze awareness," *Perception and Psychophysics*, vol. 62, no. 3, pp. 586–595, 2000.

[3] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "Human neural systems for face recognition and social communication," *Bioligical Psychiatry*, vol. 51, no. 1, pp. 59–67, 2002.

[4] S. P. Lee, Jeremy B. Badler, and Norman I. Badler, "Eyes alive," in *Proc. SIGGRAPH*. 2002, pp. 637–644, ACM Press.

[5] J. Heinzmann and A. Zelinsky, "3-D facial pose and gaze point estimation using a robust real-time tracking paradigm," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 142–147.

[6] C. Peters and C. O'Sullivan, "Bottom-up visual attention for virtual human animation," in *Computer Animation and Social Agents 2003*, 2003, pp. 111–117.

[7] D. Terzopoulos and T. F. Rabie, "Animat vision: Active vision in artificial animals," *Videre: Journal of Computer Vision Research*, vol. 1, no. 1, pp. 2–19, 1997.

[8] L G Ungerleider and M Mishkin, "Two cortical visual systems," in *Analysis of visual behavior*, D G Ingle, M A A Goodale, and R J W Mansfield, Eds., pp. 549–586. MIT Press, Cambridge, MA, 1982.

[9] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[11] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10-12, pp. 1489–1506, May 2000.

[12] E. Bizzi, "The coordination of eye-head movements," *Scientific American*, vol. 231, no. 4, pp. 100–106, 1974.

[13] D. L. Sparks, "The brainstem control of saccadic eye movements," *Nature Reviews Neuroscience*, vol. 3, no. 12, pp. 952–964, 2002.

[14] E. G. Freedman and D. L. Sparks, "Activity of cells in the deeper layers of the superior colliculus of the rhesus monkey: evidence for a gaze displacement command," *J Neurophysiol*, vol. 78, no. 3, pp. 1669–1690, 1997.

[15] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[16] A. Thiele, P. Henning, M. Kubischik, and K. P. Hoffmann, "Neural mechanisms of saccadic suppression," *Science*, vol. 295, no. 5564, pp. 2460–2462, 2002.

[17] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin, "Synthesizing realistic facial expressions from photographs.," in *Proc. SIGGRAPH*. 1998, pp. 75–84, ACM Press.

[18] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, Oct 2004.

[19] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, Jan 2005.