Joint Virtual Reality Conference of EuroVR - EGVE (2011)
R. Blach, S. Coquillart, M. D'Cruz, A. Steed, and G. Welch (Editors)

Short Paper: Exploring the Object Relevance of a Gaze Animation Model

O. Oyekoya[1], A. Steed[1] and X. Pan[1]

[1]Virtual Environments & Computer Graphics, University College London, UK

Abstract

Models for animating the eyes of virtual characters often focus on making the face appear natural and believable. There has been relatively little work in computer graphics that investigates the relevance of the objects of interest (gaze targets). In this paper, a gaze animation model has been constructed that allocates visual attention to relevant targets from objects that are within the virtual character's field of view in an immersive 3D virtual environment. Relevance is determined by proximity, eccentricity, changes in orientation and velocity of objects in the virtual character's environment. In this paper, two tasks were designed to test the relevance of the objects selected by the gaze animation model during the tasks. Eye tracking data obtained from six human subjects provided benchmark data for measuring the efficiency of the model in picking relevant objects. The gaze animation model largely outperformed a random selection algorithm in predicting the real targets/objects of users' interests within the virtual environment.

Categories and Subject Descriptors (according to ACM CCS): Computer Graphics [I.3.7]: Three-Dimensional Graphics and Realism—Virtual Reality; Computer Graphics [I.3.7]: Three-Dimensional Graphics and Realism—Animation; Simulation and Modeling [I.6.8]: Types of Simulation—Animation

1. Introduction

Autonomous Virtual Characters (AVCs) have to know when and how to interact with the objects and other characters in the virtual environment around them. The contents of their surrounding environment should compete for the AVC's visual attention, just as happens for humans in the real world. Gaze animation models have been developed for the generation of naturalistic eye movement for virtual characters but models are needed that allocate eye targets to relevant objects/characters within the virtual environment. In order to build autonomous virtual characters, we must better understand how the eyes perceive and react to stimuli in its environment.

Itti et al uses machine vision to locate targets of interest in virtual or real scenes [IDP03]. They developed a computational model that predicts the spatiotemporal deployment of gaze onto any incoming visual scene. The avatar animation model is based on the neurobiology of attention but the use of image processing algorithms to determine regions of interest makes it computationally expensive. To avoid such computational expense, we extract readily available information in computer graphics environments (i.e. position and orientation) and propose four top-down and bottom-up components of visual attention (proximity, eccentricity, velocity and orientation) to determine objects of interest, as opposed to regions of interest. Grillon et al [GT09] used similar criteria to simulate gaze behaviours for crowds but the model lacked experimental validation.

The evaluation of previous gaze models [LBB02, QBM08, MD09, MH07] have been on believability and realism while object relevance has been largely ignored. This research addresses that drawback by comparing the objects of interest computed by the gaze animation model with actual human gaze data collected with an eye tracker. Crucially, our model was constructed and evaluated in an immersive 3D virtual environment enabling a more naturalistic body control. Lee et al [LKC08] and Hillaire et al [HLRC*10] used a similar evaluation method to evaluate their respective models in a desktop virtual environment.

Yarbus' work [Yar67] demonstrated that scan-path characteristics such as their order of progression can be task dependent. Therefore this paper concerns itself with comparing human gaze behaviours with a gaze animation model [OSS09] during free-viewing and goal-oriented tasks. The

task has been designed to test the ability of the gaze animation model in detecting relevant objects within an immersive 3D virtual environment. To achieve this, a strategy was employed which allowed the effectiveness of the model to be explored further by comparing with a simple random selection algorithm. This strategy provided a performance baseline which a more intelligent approach would need to exceed. Section 2 describes the algorithms used in the paper. Section 3 presents the experimental evaluation of the model, while sections 4 presents the results and discussion. Finally, section 5 presents the conclusion.

2. Background Work

2.1. Gaze Animation Model

The gaze animation model is designed to adapt to the complex interaction within the scene. It considers varying avatar behaviour and components of visual attention (i.e. properties of objects within the scene) to compute saliency scores for all items within the field of view. The main input to this model is the virtual reality database, which stores all the objects within the scene. The model determines the target object by examining four criteria of the objects within a field of view. The horizontal field of view (fov) is set to 70° for the eye (i.e. a maximum angle of 35° towards the left or right) while the vertical fov is set to 50° (i.e. a maximum angle of 25° upwards or downwards).

1. Given the user's eye, $E = (e_x, e_y, e_z)$, and the object, $O_i = (o_x, o_y, o_z)$, the *proximity, p* is computed from the euclidean distance between the two 3D points as:

$$p = \sqrt{(e_x - o_x)^2 + (e_y - o_y)^2 + (e_z - o_z)^2}, \quad (1)$$

   A gaussian curve fit of the proximity is computed from:

$$y = f(x) = \sum_{i=1}^{n} a_i e^{\left[-\left(\frac{x-b_i}{c_i}\right)^2\right]}, \quad (2)$$

   The Gaussian model is used for fitting peak and was generated from eye tracking data of twelve users [OSS09]. It is given by the equation 2 where a_i are the peak amplitudes, b_i are the peak centroids (locations), and c_i are related to the peak widths, n is the number of peaks to fit, and $1 \leq n \leq 8$. Proximity, p is fitted with the values $a_1 = 19.11$, $a_2 = 6.68$, $b_1 = 1.83$, $b_2 = 3.27$, $c_1 = 0.87$ *and* $c_2 = 1.7$. The saliency score, S_p of the object's proximity is also computed from equation 2 where $x = p$ and is normalised by dividing by a_1 (i.e. peak amplitude) to keep the range between 0 and 1.
2. The *eccentricity,* θ defined as the magnitude of the dot product is computed as:

$$\theta = \arccos\left(\frac{u \cdot v}{|u||v|}\right), \quad (3)$$

   where $u = (u_x, u_y, u_z)$ is the head-centric vector and $v = (v_x, v_y, v_z)$ is the direction vector of the eye to the object, $(e_x, e_y, e_z) - (o_x, o_y, o_z)$. A gaussian curve fit of the eccentricity is computed from equation 2. Eccentricity, θ is fitted with the values $a_1 = 40.13$, $a_2 = 8.09$, $b_1 = 14.39$, $b_2 = -14.05$, $c_1 = 4.18$ *and* $c_2 = 40.5$. The saliency score, S_θ of the object's eccentricity is computed from equation 2 where $x = \theta$ and is normalised by dividing by a_1 (i.e. peak amplitude).
3. *velocity, v* is defined as the rate of change of the object's location and is computed as:

$$v = \frac{\Delta O_i}{\Delta t}, \quad (4)$$

   where ΔO_i is the euclidean distance between an object's location at time t_1 and its location at time t_2, and Δt is the time interval of the frame duration. The normalised saliency score, S_v of the object's velocity is given by $v/20$ (i.e. a reasonable maximum speed of 20 feet per second).
4. *orientation,* Δq defined as the change in object's angular position over time and is computed as:

$$\Delta q = 2 \arccos({q_1}^{-1}.q_2) \quad (5)$$

   where quaternions q_1 and q_2 represent two orientations at time t_1 and t_2 respectively. The normalised saliency score, $S_{\Delta q}$ of the object's orientation is given by $\Delta q/180$ (i.e. a reasonable maximum change in orientation of 180°).

The saliency of each object within the field of view is computed from a summation of the normalised saliency scores.

$$S_O = S_\theta + S_p + S_v + S_{\Delta q}, \quad (6)$$

The summation S_O is then used to guide attention, as the selected target in each frame will become the object with the highest overall saliency score. A fixation towards the particular object will then occur. The fixation duration for the eye is limited to 300ms as long as the target object remains within the field of view (according to Henderson's average duration during scene viewing [HH99]).

The eyeball is interpolated over 6 frames by fitting to an exponential velocity curve as presented in Lee et al [LBB02], [VGSS04]

$$y = 14e^{[-\pi/4(x-3)^2]}, \quad (7)$$

where $x = frame\{1,2,3,4,5,6\}$. The eye is moved to intermediate positions within each frame to produce a smooth movement during saccades.

In order to decrease the probability of the model continually choosing centered objects, the random selection algorithm described in section 2.2 computes the target object on 25% of the time while the gaze animation model computes the target object on 75% of occasions. It also means that the eye is animated even when the virtual character is idle. However this introduces a level of unpredictability into the model, necessitating further scrutiny of the model's predictive capabilities in this paper.

2.2. Random Selection Algorithm

The random algorithm determines the target object by picking randomly from the objects within the field of view (same as above). Saccades and fixations are randomly distributed between objects within the current field of view. Thus, as users move their heads, potential targets enter and exit the field of view, and new saccades and fixations will be generated. Fixation duration on the objects of interest are determined by a random sampling method which is varied by the head motion.

3. Experiment

In order to determine the accuracy of the eye-gaze model in picking relevant objects, the gaze model and the random selection algorithm are compared to a benchmark data gathered from tracked gaze.

1. Tracked gaze: head mounted mobile eye trackers were worn by each participant as they performed the task.
2. Gaze Model: as described in section 2.1.
3. Random Selection Algorithm: as described in section 2.2.

We hypothesize that the tracked gaze data will match more closely with the discriminatory gaze model data when compared to the non-discriminatory random data. Performance was measured in two scenarios (Figure 1):

1. Goal-oriented task: An object manipulation task (i.e. solving a cubes puzzle) was designed. In this scenario, participants were instructed to solve a puzzle on their own involving eight cubes with various colours on each side, and to arrange them into a larger cube such that each side would display exactly one single colour.
2. Free-viewing task: A town navigation scenario (walking through a large town scene) was desinged and the same participants were instructed to explore the town environment without any specific goal in mind.

3.1. Data Collection

Data was collected from six naïve participants performing the two tasks. Both tasks took an average of 5 minutes each. The eye behaviour of the users was captured within an immersive virtual environment platform (CAVE™-like system) [WRM*08] operating at 60fps. During both sessions, the user wore a head tracker and held one hand tracker. The user was also calibrated with a head-mounted mobile eye tracker to drive avatar gaze in real-time. Binocular eye-trackers from Arrington Research, Inc. were mounted on the CAVE's CrystalEyes®3 shutter-glasses. Log files were recorded for each of the two virtual environment scenarios described above. The hand tracker was used as an input device in the puzzle scene to manipulate objects while the joystick on the hand tracker was used to navigate around the town. The logfiles recorded the gaze targets for the tracked gaze, gaze animation model and the random selection algorithm.

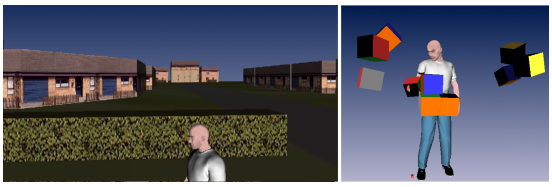


Figure 1: *Town Navigation and Cubes Puzzle Scenes.*

Given the participants' mapped eye and head position, ray casting or surface intersection tests enabled an accurate estimation of the object of interest for the tracked gaze in each frame. As participants performed the task wearing the eye tracker, two seperate background processes computes the gaze targets for the gaze animation model and random selection algorithm respectively. This enabled comparison of the gaze targets in terms of the relevance of the prediction and the duration of fixations during the session.

4. Results and Discussion

In the town navigation scenario where there were 147 possible targets with an average spacing between targets of 200.38 units and average size of 49.79 units. The cubes puzzle scenario had 78 possible targets (13 cubes with 6 surfaces each) with an average spacing between targets of 2.94 units and an average size of 1.18 units. Each scene had different properties: targets were larger and more spaced out in the navigation scenario unlike the cubes puzzle scenario.

In order to assess the relevance of the model's prediction, the real gaze targets of the tracked gaze for each participant was compared with the gaze targets computed by the gaze animation model and the random selection algorithm respectively. Figure 2 indicate the extent (as a percentage proportion of the task duration) to which the eye tracker and the gaze animation model produced gaze at the same target objects in the environment, compared with the extent to which the eye tracker and the random selection algorithm produced gaze at same targets. Generally, the gaze animation model mostly outperforms the random selection algorithm except on two sessions in the goal-oriented cubes puzzle task and one session in the free-viewing town navigation task. It must be pointed out that the percentages are very low, less than 10%. The problem is that we can't know the particular visual strategy of the participants. However, the random selection algorithm provided a performance baseline that the gaze animation model needed to exceed. Figure 2 (navigation) in particular shows significant improvement in the time spent looking at the same objects as the user. It is particularly notable from a subjective point of view that the random selection algorithm is simply implausible in this case.

A paired t-test on the proportion of time that model's target prediction equals tracked gaze target shows that in the Town scenario, the gaze model had a better performance than random ($p<0.04$). However there was no significant difference between the gaze model and the random one

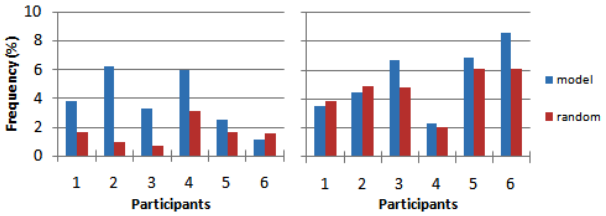
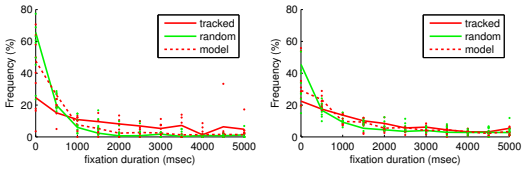


Figure 2: *Proportion of time that model's target prediction equals tracked gaze target. Left:Town Navigation (free-viewing task). Right: Cubes Puzzle (goal-oriented task).*



(a) town (b) cubes

Figure 3: *Comparisons of fixation duration*

(p=0.18) in the Cube scenario. A test of the residuals of the models does not reject the hypothesis of normality. The gaze model's prediction clearly outperformed the random algorithm's prediction in the free-viewing town navigation scenario and this is likely attributed to a poorer performance of the random selection algorithm when compared with the goal-oriented cubes puzzle scenario. The wider spacing of the targets and increased movements due to the navigation may also have increased the chance that irrelevant random gaze targets are selected while the character is on the move.

Figures 3(b) and 3(a) show a comparison of the fixation durations along with the spread for all participants. The 500msecs peaks in the fixation durations of the random selection algorithm was evident in both the free-viewing and goal-oriented tasks. A Jarque-Bera test rejected the normality of the tracked ($p<0.01$) and the random data ($p=0.02$), however using the log value of the data eliminated this problem. A paired t-test rejects the equality between the log value of the random and the tracked data ($p<0.01$) but not between the log value of the model and the tracked data ($p=0.89$). This shows that the random data is significantly different from the tracked data, but there is no statistical significance between the model and the tracked data. On average, the gaze model's peak is reduced but it clearly shows that the fixation duration can be tuned and improved further to match with tracked gaze data.

5. Conclusion

Although the gaze animation model was not built specifically for a specific environment, task or user, this study has shown that the model generally seems to try to adapt reasonably well in explorative mode. The results identify areas for further improvements, particularly goal-oriented tasks. The question of how to weight the saliency criteria for a better performance remains a matter for further research. Indeed the spread of the fixation durations clearly need to be adjusted by varying the fixation duration threshold rather than the current limit of 300ms.

The improved performance of the gaze animation model relative to the random selection algorithm in the free-viewing task highlights a promising area to explore in creating believable virtual characters. Future work will concentrate on including more parameters (such as shape or size of objects), adapting the model to drive the head and automatic navigation of a virtual character. This research moves us towards the goal of building autonomous virtual characters that know when to interact with objects and other virtual characters in the environment around them.

References

[GT09] GRILLON H., THALMANN D.: Simulating gaze attention behaviors for crowds. *Computer Animation and Virtual Worlds, 20 2*, 3 (2009), 111–119. 1

[HH99] HENDERSON J., HOLLINGWORTH A.: High-level scene perception. *Annual Review of Psychology* (1999), 243–244. 2

[HLRC*10] HILLAIRE S., LÉCUYER A., REGIA-CORTE T., COZOT R., ROYAN J., BRETON G.: A real-time visual attention model for predicting gaze point during first-person exploration of virtual environments. In *Proc. of 17th ACM VRST* (2010), pp. 191–198. 1

[IDP03] ITTI L., DHAVALE N., PIGHIN F.: Realistic avatar eye and head animation using a neurobiological model of visual attention, 2003. 1

[LBB02] LEE S., BADLER J., BADLER N.: Eyes alive. *ACM Transactions on Graphics 21*, 3 (2002), 637–644. 1, 2

[LKC08] LEE S., KIM G., CHOI S.: Real-time tracking of visually attended objects in virtual environments and its application to lod. *IEEE Transactions on Visualization and Computer Graphics* (2008), 6–19. 1

[MD09] MA X., DENG Z.: Natural Eye Motion Synthesis by Modeling Gaze-Head Coupling. In *Proc. of IEEE Virtual Reality Conference* (2009), pp. 143–150. 1

[MH07] MASUKO S., HOSHINO J.: Head-eye Animation Corresponding to a Conversation for CG Characters. In *Computer Graphics Forum* (2007), vol. 26, Blackwell Synergy, pp. 303–312. 1

[OSS09] OYEKOYA O., STEPTOE W., STEED A.: A saliency-based method of simulating visual attention in virtual scenes. In *Proc. of 16th ACM VRST* (2009), pp. 199–206. 1, 2

[QBM08] QUEIROZ R., BARROS L., MUSSE S.: Providing expressive gaze to virtual animated characters in interactive applications. *Computers in Entertainment (CIE) 6*, 3 (2008). 1

[VGSS04] VINAYAGAMOORTHY V., GARAU M., STEED A., SLATER M.: An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. In *Computer Graphics Forum* (2004), vol. 23, John Wiley & Sons, pp. 1–11. 2

[WRM*08] WOLFF R., ROBERTS D., MURGIA A., MURRAY N., RAE J., STEPTOE W., STEED A., SHARKEY P.: Communicating Eye Gaze across a Distance without Rooting Participants to the Spot. In *Proc. IEEE/ACM DS-RT 2008* (2008), pp. 111–118. 3

[Yar67] YARBUS A.: *Eye movements and vision*. Plenum press, 1967. 1