A Saliency-Based Method of Simulating Visual Attention in Virtual Scenes

Oyewole Oyekoya[*] William Steptoe[†] Anthony Steed[‡]
University College London University College London University College London

Abstract

Complex interactions occur in virtual reality systems, requiring the modelling of next-generation attention models to obtain believable virtual human animations. This paper presents a saliency model that is neither domain nor task specific, which is used to animate the gaze of virtual characters. A critical question is addressed: What types of saliency attract attention in virtual environments and how can they be weighted to drive an avatar's gaze? Saliency effects were measured as a function of their total frequency. Scores were then generated for each object in the field of view within each frame to determine the most salient object within the virtual environment. This paper compares the resulting *saliency gaze model* to *tracked gaze*, in which avatars' eyes are controlled by head-mounted mobile eye-trackers worn by human subjects, *random gaze model* informed by head-orientation for saccade generation, and *static gaze* featuring non-moving centered eyes. Results from the evaluation experiment and graphical analysis demonstrate a promising saliency gaze model that is not just believable and realistic but also target-relevant and adaptable to varying tasks. Furthermore, the saliency model does not use any prior knowledge of the content or description of the virtual scene.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation I.6.8 [Simulation and Modeling]: Types of Simulation—Animation

Keywords: Character animation, gaze modeling, visual attention, target saliency, behavioural realism, facial animation

1 Introduction

This research aims to construct gaze models for virtual characters that mimic real human eyes. In virtual scenes, salient stimuli such as the changing location and orientation of objects attract attention providing a method of allocating attention to objects within virtual environments. In order to mimic real human eyes, a better understanding of how humans allocate visual attention is needed. Humans cannot attend to all things at once, thus our attention capability is used to focus our vision on selected regions of interest. Our capacity for information processing is limited, therefore visual scene inspection is performed with particular attention to selected stimuli of interest. A good definition of visual attention was given by James [James 1890]: *"Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. ... It implies withdrawal from some things in order to deal effectively with others"*. This definition implies that visual attention modelling is relevant to the objective of obtaining better scene content understanding as employed in neurobiological models of visual attention [Itti et al. 1998; Stentiford 2007]. In natural scenes, attentional selections are influenced by complex interactions between top down, goal driven control and bottom up, stimulus driven control [Itti et al. 1998]. Understanding of how visual targets compete for attention is critical in understanding how gaze is allocated to targets within a visual scene. While attention has been shown to be influenced heavily by top-down components refers to visual inspection with a task in mind [Yarbus 1967] (e.g., looking for John Doe in a crowd), it has also been shown to be strongly influenced from the bottom-up by the contents of the visual input enabling the development of computational models. One such model [Itti et al. 1998] has been used to synthesize realistic gaze motions in avatars [Itti et al. 2003] but it is unclear how the model will perform in a virtual scene where colour and texture features are deliberately tweaked to be unnatural (e.g. Arts). Furthermore, the authors admit that a strong limitation of their approach is its computational cost and suggest that it may be promising to leverage information from virtual scenes to detect salient features with little or no image processing. This research draws support from that suggestion.

Virtual scenes are digital 3D representations of natural scenes and contain a set of objects, such as items and avatars that move and interact in 3D space. Virtual reality scenes make use of vector graphics such as shapes or polygons to represent objects in computer graphics. These objects can have *intrinsic saliency* in terms of their *proximity*, *eccentricity*, *orientation* and *velocity* while their *extrinsic saliency* is given by its interest to the avatar such as the *fixation duration*. *Proximity* relates to the euclidean distance between the user's eye from the object. *Eccentricity* is based on the angular distance of objects from the center of gaze (head-centric vector) allocating attention to objects under direct scrutiny than objects in the user's peripheral vision. *Velocity* is based on the object's speed across the user's visual field allocating attention to objects moving quickly across the user's gaze than slow-moving or still objects. *Orientation* is based on objects' differences in rotation with attention allocated to objects with higher rotation speed. Saliency can thus be inferred from the user's actions such as a waving avatar hand (orientation), a moving car (velocity) or a close object (proximity) under scrutiny (eccentricity). Previous work has used some of these saliency parameters to vary the level of detail rendering of an object [Luebke 2003] and to simulate gaze attention behaviours for crowd animations [Grillon and Thalmann 2009]. In this paper, the intrinsic saliency of an object is computed from the location and/or orientation of the object in 3D space. The following section reviews the literature on previous gaze models and approaches to modelling attention. Section 3 describes initial work done to obtain the data used to construct the gaze model. Section 4 presents a technical description of the saliency gaze model. Section 5 presents the evaluation of the model, while sections 6 and 7 presents the discussion and conclusions respectively.

2 Related Work

2.1 Previous Gaze Models

Gaze models have been developed for the generation of naturalistic eye-movement for virtual characters. Previous studies have inves-

[*] e-mail: W.Oyekoya@cs.ucl.ac.uk
[†] e-mail:W.Steptoe@cs.ucl.ac.uk
[‡] e-mail:A.Steed@cs.ucl.ac.uk

tigated dyadic conversation using a gaze model that is informed by interactional states (*e.g.* speaking and listening) [Lee et al. 2002; Vinayagamoorthy et al. 2004] and has been found to significantly improve the perceived quality of avatar-mediated communication [Garau et al. 2003]. Most eye gaze models have been based on extensive psychological studies of the functions of gaze behaviour such as the exchange of social signals [Argyle and Cook 1976]. Hence gaze patterns are associated with certain cognitive states. Lee et al [Lee et al. 2002] used an eye saccade statistical model during talking and listening based on empirical eye tracking data. These values implement statistical generalisations about human gaze behaviour derived from empirical studies of saccades and/or statistical models of eye-tracking data. Peters et al [Peters et al. 2005] presents a model for a virtual character (referred to as embodied conversation agent) that is able to start, maintain and end a conversation based on the level of interest and engagement of the other virtual character. The algorithm integrates the listening and speaking state into a model able to sustain a conversation between two virtual characters by monitoring the level of interest of each character. Gaze models can be driven by cognitive operations [Khullar and Badler 2001] and generates gaze behaviour that reflects the agent's inner thoughts, including continuous gaze following and gaze aversion [Lee et al. 2007]. Queiroz and Barros models expressive gaze by examining eye behaviour in different affective states from computer graphics movies [Queiroz et al. 2008]. Eye motion has also been generated given a head motion sequence as input, by statistically modelling the coupling between gaze and head movement [Ma and Deng 2009], in addition to synchronization with content of utterance and state of conversation [Masuko and Hoshino 2007].

Current gaze models suffer from three drawbacks: (i) assumptions are made about the gaze patterns that relate to certain social signals and cognitive state with limited understanding of how they fit into a temporal dimension, (ii) detecting cognitive states tends to add another layer of input which is not readily available in computer graphics systems, (iii) the focus of these gaze models have been on believability and realism while target relevancy has been largely ignored. These models tend to be tested in scenarios where the avatar is engaged in a conversation with a non-moving target that is straight ahead or in a multi-party conversation where there are two or more other avatars as targets [Gu and Badler 2006]. This research is concerned with single or multi-party interaction where targets can be other avatars and lifeless objects alike.

2.2 Attention Modelling

The approach used in this paper draws support from previous research in visual attention modelling on static images and dynamic video scenes. Based on the feature integration theory [Treisman and Gelade 1980] derived from visual search experiments, Koch and Ullmans framework [Koch and Ullman 1985] for simulating human visual attention focuses on the idea that the control structure underlying visual attention needs to represent such locations within a topographic saliency map, especially given that the purpose of visual attention is to focus computational resources on a specific, conspicuous or salient region within a scene. Multiple image features such as colour, orientation and intensity are combined to form a saliency map that reflects areas of attention. In the same way, an object's intrinsic saliency (defined in [Findlay and Walker 1999]) can be derived from parameters such as its proximity, eccentricity, orientation and velocity of the objects in a virtual reality scene. Computation of the intrinsic saliency determines the spatial coding of fixations in the virtual scene. As in research on images, the question of which saliency parameter is more important at any point in time is dependent on the spatial clustering of objects and the complexity of interaction in the virtual scene.

The extrinsic saliency of an object determines the duration of fixations. It is concerned with coherent fixation distributions during inspection of a virtual scene. Henderson and Hollingworth [Henderson and Hollingworth 1999] review this area of high-level scene perception research further, which concerns the role of eye movements in scene perception, focusing on the influence of ongoing cognitive processing on the position and duration of fixations in a scene. They speculate on whether ongoing perceptual and semantic processing accounts for the variability of fixation durations, which range from less than 50ms to more than 1000ms in a skewed distribution with a mode of about 230ms. The average fixation duration during scene viewing is also said to be 330ms, with a significant variability around this mean. Their review of eye movement studies during scene viewing suggests that fixation positions are nonrandom, with fixations clustering on both visually and semantically informative regions. They also found that the spatial distribution of the first few fixations in a scene seems to be controlled by the visual features in the scene and the global (not local) semantic characteristics of the scene. As viewing progresses and local regions are fixated and semantically analyzed, positions of later fixations come to be controlled by both the visual and semantic characteristics of those local regions. The length of time the eyes remain in a given region is immediately affected by both characteristics. Presented in this paper is a plausible model driven by the controlling user's head orientation, which is used to determine the extrinsic saliency (i.e. fixation duration) of an object in a virtual reality scene.

Previous research leads to the hypothesis that the eye is attracted to regions of a virtual scene that convey what is thought at the time to be the most important information for scene interpretation. The intrinsic saliency of an object in a virtual scene determines the spatial distribution of fixations, inferred from continuous interaction within the virtual scene. The extrinsic saliency drives the temporal coding of fixations (i.e. duration), thus presenting a plausible and coherent saliency gaze model. The approach also implements a plausible linear interpolation algorithm for the dynamics of the eyeball.

3 Background Work

An earlier study [Steptoe et al. 2009] was conducted to evaluate three methods of avatar eye-gaze control (*tracked gaze*, *random model* and *static gaze*) during an object-focused puzzle scenario performed between three networked Immersive Collaborative Virtual Environment (ICVE) systems. Twelve participants took part in this study in a repeated measures experiment where they were represented by avatars in a shared 3D space. Participants were able to move freely and manipulate virtual objects within the ICVE enabling the collection of detailed logs of participants' behaviour. Head mounted mobile eye trackers were worn by each participant which provided the *tracked gaze* data. Hence this system provides an ideal platform for investigating gaze behaviour in virtual scenes. In the second condition, the *random model* (described below) was used to drive the avatar's gaze while they conducted the task. The ICVE platform is built on OpenSG®, an open source scenegraph system for interactive 3D graphics applications that contains a collection of nodes in a graph or tree structure with a parent-child architecture. Each object within the object puzzled scene is represented by a node, associated to a geometrical transformation matrix that contains location and orientation data at any time in the virtual scene.

3.1 Random Gaze Model

From the outset, the main input to the random model is the scene database, which stores all the objects within the virtual reality scene. Algorithm 1 determines the target object by picking randomly from the objects within the field of view. The field of view, fov is set to 70°, regarded as 35° eccentricity (computed from equation 1). Saccades and fixations are randomly distributed between targets 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. By seeding the random number generator to change every 1 second, a uniformly distributed random number is generated every second. The temporal coding of fixations is therefore dependent on timing and velocity of head movement: reduced activity generates fewer saccades with longer fixations, while rapid motion results in greater numbers of saccades with shorter fixation times. The random gaze model is thus informed by a user's current field of view inferred from head orientation to generate eye gaze animation throughout an unfolding interaction.

Algorithm 1 *random model* computes target object (o_x, o_y, o_z)

```
Require: scene database of objects ${O_1, O_2, O_3, ...O_n}$
Require: field of view, $fov = 70°$ (i.e. eccentricity, $\theta \leq 35°$)
for each frame do
    seed random to 1 second {determines fixation duration}
    compute avatar's eye location in world coordinates
    for each object in the scene database do
        determine object's location in world coordinates
        compute eccentricity, $\theta$ {equation 2}
        compute vertical angle, $\theta_v$ {equation 3}
        if $(\theta < 35°)$ and $(-25° < \theta_v < 25°)$ then
            add to list of objects within field of view
        end if
    end for
    pick randomly from objects within field of view
    aim avatar's eyes at center of selected target object
end for
```

3.2 Critique of the Random Model

The frequency plots in Figure 1 shows the combined frequencies of five gaze behaviours (proximity, saccade magnitude, saccade velocity, fixation duration and the eccentricity) from all twelve participants. A comparison of tracked gaze with the random gaze model showed a clear difference between the plots for each gaze parameter. The spread of the eccentricity on the random model demonstrates the randomness of the targets chosen, as compared to the peakedness of the tracked gaze. The fixation duration plot for the random gaze model against tracked gaze shows a peak in the random gaze model's fixation durations at the 500ms mark demonstrating why the model largely underperformed. Therefore, the goal is to construct a saliency gaze model with a highly-correlated plot to tracked gaze. A gaussian curve fit of the proximity and eccentricity of the tracked gaze data is computed from:

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

The Gaussian model is used for fitting peaks, and is given by the equation 1 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* $c_2 = 1.7$. 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* $c_2 = 40.5$.

4 Saliency Gaze Model

The saliency model is designed to adapt to the complex interaction within the scene. It considers varying avatar behaviour and properties of objects within the scene. Saliency computation is based on the likelihood functions generated from section 3.2. The distribution of saliency values and objects of interests are often spatially biased towards the center of the view point [Melcher and Kowler 2001]. In order to decrease the probability of center bias in the saliency model, the random model computes the target object on 25% of the time while the saliency model computes the target object on 75% of occasions. To implement this, a uniformly distributed random number 0 and 3 is generated and the random model is implemented instead whenever the random number 3 is generated (i.e. when saliency state equals false as described on line 23 to 27 of algorithm 2. The probability that the saliency algorithm is used is thus given by: $P(salience) = 3/4$.

4.1 Spatial and Temporal Distribution of Fixations

The main input to this model is the scene database, which stores all the objects within the scene. Algorithm 2 determines the target object by examining the intrinsic saliences of the objects within the field of view (35° eccentricity).

Algorithm 2 *saliency model* computes target object (o_x, o_y, o_z)

```
Require: scene database of objects ${O_1, O_2, O_3, ...O_n}$
Require: field of view, $fov = 70°$ (i.e. eccentricity, $\theta \leq 35°$)
 1: for each frame do
 2:     include phantom object for head-centric vector in scene database
 3:     compute elapsed time since previous frame $\Delta t$
 4:     seed random to 1 second {determines saliency state and fixation duration}
 5:     compute avatar's eye location in world coordinates $(e_x, e_y, e_z)$
 6:     for each object in the scene database do
 7:         determine object's location and orientation in world coordinates
 8:         compute eccentricity, $\theta$ {equation 3}
 9:         compute vertical angle, $\theta_v$
10:         if $(\theta < 35°)$ and $(-25° < \theta_v < 25°)$ then
11:             compute saliency scores for:
12:             - change in orientation, $\Delta q$ {equation 5}
13:             - object's velocity $v$ {equation 4}
14:             - proximity $p$ {equations 1 & 2}
15:             - eccentricity $\theta$ {equations 1 & 3}
16:             compute total saliency score and store in list, A
17:             add to list of objects within field of view, B
18:         end if
19:     end for
20:     if $(\Delta tc < 300ms)$ and (current target object is still within fov) then
21:         return current target object's location {section 3.1}
22:     end if
23:     if saliency state = true then
24:         determine object with highest saliency from list A
25:     else {considers less salient objects (in the periphery)}
26:         pick randomly from objects within field of view from list B
27:     end if
28:     aim avatar's eyes at center of selected target object
29:     if previous target $\neq$ selected target then
30:         compute elapsed time since last target changed, $\Delta tc$
31:         compute eyeball interpolation {equation 7}
32:     end if
33: end for
```

4.1.1 Intrinsic Saliency Criteria

The saliency gaze model generates eye target based on four intrinsic saliency of the objects:

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:

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

The saliency score, S_p of the object's proximity is also computed from equation 1 where $x = p$ and is normalised by di-

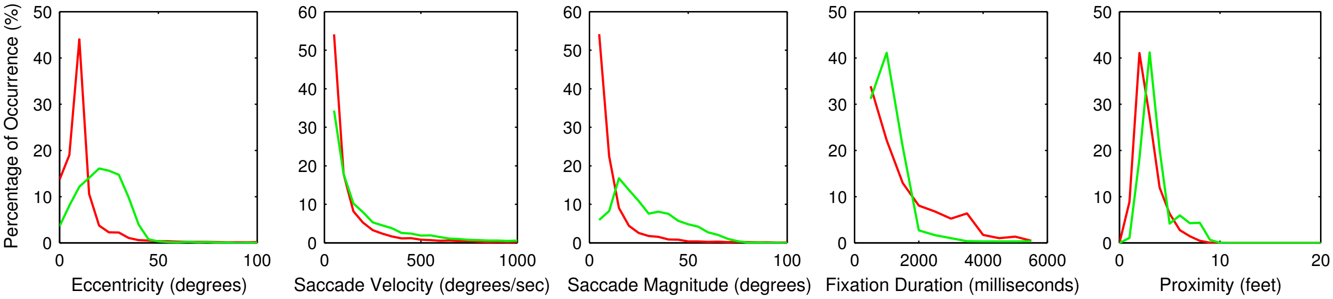


Figure 1: *Comparisons of Gaze Parameters (— tracked, — random).*

viding 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:

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

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, $(e_x, e_y, e_z) - (o_x, o_y, o_z)$. The saliency score, S_θ of the object's eccentricity is computed from equation 1 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:

$$v = \frac{\Delta O_i}{\Delta t}, \tag{4}$$

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:

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

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°).

4.1.2 Saliency Scoring and Fixation Duration

The saliency of each object within the field of view is computed from a summation of the normalised saliency scores and is used to guide attention.

$$S_O = S_\theta + S_p + S_v + S_{\Delta q}, \tag{6}$$

The object with the highest combined saliency score is determined as the target. The computation of these scores relies on appropriate normalisation and summation steps in a competitive way to determine most likely target object to be allocated fixations. The fixation duration is limited to 300ms as long as the target object remains within the field of view (in line with Henderson's average duration during scene viewing [Henderson and Hollingworth 1999]).

4.2 Eyeball Dynamics

The eyeball is interpolated over 6 frames by fitting to an exponential velocity curve as presented in Lee et al [Lee et al. 2002; Vinayagamoorthy et al. 2004].

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

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.

5 Experiment

The following experiment evaluates four methods of eye gaze control across three different scenarios. The four methods of gaze control were as follows:

1. *None*: static, centred eyes (N).
2. *Random gaze model*: as described in section 3.1 (R).
3. *Saliency gaze model*: as described in section 4 (S).
4. *Tracked gaze*: reproduction of recorded gaze from an eye tracker (T).

When operating under condition N (static gaze), the avatar's eyes remain focussed directly ahead. In this condition, head-orientation is relied upon as the primary indicator of visual attention. In condition T (tracked gaze), the avatar's eyes reproduce a wearer's eye movements.

In order to assess the ability of the saliency model to generate natural and meaningful gaze, it was important to measure performance in varying scenarios. Therefore, three virtual environments were designed (figure 2), which aimed to place users in three practical situations: object manipulation and design (solving a puzzle), telecommunication (engaging in two-party conversation), and navigation (walking through a large town scene).

5.1 Data Collection

The performance of an expert user of the ICVE system was captured within a collaborative virtual environment platform [Wolff et al. 2008] operating at 60fps in an immersive CAVE$^{\text{TM}}$-like system. Sessions were captured each of the three virtual environments described above and illustrated in figure 2. During the object-focused puzzle scenario and the navigation scene, the system operated in single-user mode, while the system operated in multi-user mode during the conversational scenario, with an avatar representing each user engaging in an informal conversation. The user performed each scene under the four eye gaze control conditions (N,R,S,T), and engaged in the puzzle, conversation, or navigation, in a relaxed and natural manner. Following data capture, a replay tool [Murgia et al. 2008] and Fraps®(Beepa®) were used to create 72 video-clips each of around 15 seconds in duration at a matching 60fps and at a resolution of 592(h)x384(v) to create the stimuli for the user study.

During all sessions, the user wore a head tracker and held one hand tracker, while in the tracked gaze (T) sessions only, the user was also calibrated with a head-mounted mobile eye tracker to drive

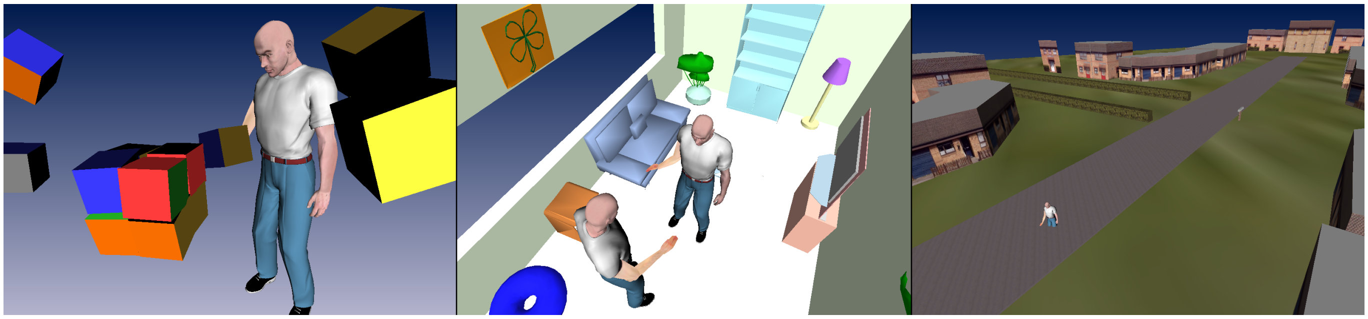
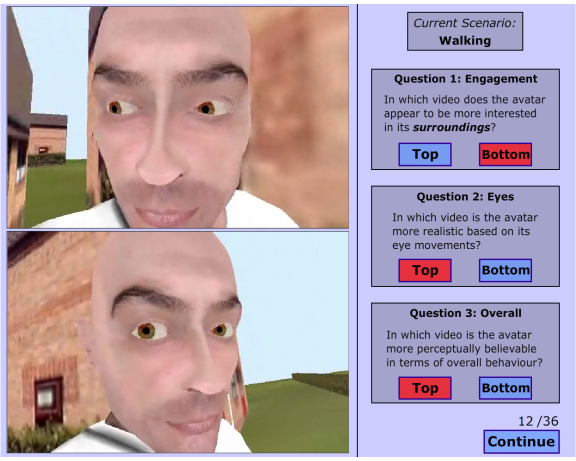


Figure 2: *Screenshots from the three virtual scenes: object-focussed puzzle (left), conversation with another user represented by an avatar (center), and navigating through a large town environment.*



Current Scenario: Walking

Question 1: Engagement

In which video does the avatar appear to be more interested in its ***surroundings***?

Question 2: Eyes

In which video is the avatar more realistic based on its eye movements?

Question 3: Overall

In which video is the avatar more perceptually believable in terms of overall behaviour?

Figure 3: *Screenshot of experiment interface. The camera was set to pan and rotate closely around the avatar's face, thus providing a varying and clear view as would be the norm in actual use.*

avatar gaze in real-time. Thus, several means were used to animate various components of the avatar. Firstly, the avatar's eyes were animated by the eye tracker (T) or by a gaze model (S,R), or were still (N). Motion of the head tracker was mapped to the avatar's head, and was also used as input to an inverse kinematic model which inferred rotation and posture of the avatar's body. Similarly, motion of the hand tracker was used to animate the avatar's right arm and hand, which was of particular use to communicate gesture and expression in the conversation scene. The hand tracker was also used as an input device in the puzzle scene to manipulate objects. As the puzzle and navigation scenarios were single-user evaluations, speech was not featured, and hence the avatar's mouth remained closed throughout the capture. In the conversation scene, mouth-movement of both avatar was animated by a speech detector and the second avatar's gaze was always driven by the saliency gaze model.

The final component of the avatar's animation were two behavioural models which generated realistic eyelid kinematics: a 'lid saccade' model initiated vertical shifts in eyelid position according to changes in vertical rotation of the eyes, and a blink model animated the rapid closing and opening of the eyes. The lid saccade model takes eye gaze data as input, and thus was driven by the eye tracker (condition T) or gaze models (conditions S and R), while it remained still in condition N. Blinks were generated every 3.5 seconds (17/minute), as the average blink rate for a person at rest as defined by [Bentivoglio et al. 1997].

5.2 Experimental Design

A balanced design paired comparison test was conducted which required subjects to judge aspects of the avatar's behaviour over a series of paired animations. The subjects were 70 volunteers (prize draw incentive) who performed the experiment online at their own machine and in their own time. While the experimental instructions provided subjects with clear guidance, we chose not to perform controlled lab-based experiments in favour of the online method's ability to encompass a range of varying displays and to enable us to reach more volunteers.

As stated, conditions were N (non-moving static eyes), R (random gaze model), S (saliency gaze model), and T (tracked gaze), thus resulting in six unique comparison pairs, and 18 over the three virtual environments. However, in order to negate the influence of vertical placement (see figure 3), the sequence was repeated with opposite top/bottom placement and randomised. Subjects were instructed to evaluate each of the 36 video pairs by answering three questions focusing on different aspects of the avatar's behavior. For each question, the subject would select *Top* if they judged the upper avatar favorably or *Bottom* if they preferred the lower avatar. Figure 3 shows the experiment interface and the questions asked. The questions were designed to extract information regarding Q1) how involved in the particular scenario (object-focussed puzzle, conversation, or navigation) the avatar appeared to be, Q2) the natural quality of eye motion, and Q3) the overall realism of the avatar. It was important to design the experiment to generate data for analysis which maintained the semantic contexts of the three virtual environment scenarios. Therefore, while always eliciting metrics of engagement, Q1's phrasing varied slightly between three versions according to the scenario currently under inspection: In which video does the avatar appear to be more (*engaged in the puzzle / engaged in the conversation / interested in its surroundings*)? This framing of Q1 also served to contextualise the following two questions, which focused on the realism of the eye motion (Q2), and the overall perception of the avatar's believability (Q3).

5.3 Analysis

After recording the participants votes on the 36 pairs, a preference matrix was computed for each of the three scenes based on the voting results of the 70 subjects. The preference matrices are shown in Tables 2(a) - puzzle scene, 2(b) - conversation scene, and 2(c) - navigation scene. The number in each cell denotes the selection frequency of a specific method when answering one of the three questions, with 1 point given for each choice. For instance, in table 2(a), '110' in the first cell of the final row indicates that condition T (tracked gaze) was voted a total of 110 times better than condition N (static gaze) in terms of overall realism (Q3). Correspondingly, the fourth cell on the first row indicates that condition N was voted a total of 30 times better than condition T for the same question. Note that the totals for each such corresponding condition pair sum to 140 (70 participants and two ratings with reversed vertical position).

Prior to performing vote comparisons for the four gaze conditions, two statistical tests were carried out, proposed by Kendall and Smith [Kendall and Smith 1940]: 1) coefficient of consistency (ζ) for each participant in order to determine whether there was any intransitive vote, and 2) coefficient of agreement (μ) to see whether the participants voted for all the pairs in a similar way. Table 1 shows the averaged coefficient of consistency for all participants for each question and scene (S1 - puzzle, S2 - conversation, S3 - navigation). The coefficient of agreement is also shown for each question and scene, together with the corresponding chi square (χ^2) value and respective significance value (p) given the six degrees of freedom. The chi square results indicate that for the used research questions and samples there is a statistically strong agreement among experimental subjects.

Table 1: *Comparisons of consistency (ζ) and agreement (μ) test statistics. Chi square (χ^2) and related* ***p*** *values given 6df. Ranking of conditions for each question and scene also shown.*

S#,Q#	ζ	μ	χ^2	p, 6 *d.f.*	1st	2nd	3rd	4th
1,1	0.380	0.079	71.9	<0.001	T	S	R	N
1,2	0.393	0.181	157.1	<0.001	T	S	R	N
1,3	0.384	0.158	137.8	<0.001	T	S	R	N
2,1	0.320	0.123	108.8	<0.001	S	T	R	N
2,2	0.386	0.186	161.4	<0.001	T	S	R	N
2,3	0.393	0.174	151.3	<0.001	T	S	R	N
3,1	0.493	0.253	217.0	<0.001	S	R	T	N
3,2	0.520	0.249	213.5	<0.001	T	S	R	N
3,3	0.459	0.237	203.4	<0.001	T	S	R	N

The results were statistically significant ($p<0.001$) between the gaze conditions. However, this does not determine between which conditions the significances lie. Hence, a series of *multiple comparison score* tests was performed as described by Ledda et al. [Ledda et al. 2005] to test the scores of the six pairs of conditions, thereby establishing where statistical differences lie. Figure 4 illustrates the significances between the gaze conditions for all scenes and questions. Conditions are ranked from lowest to highest (left to right) as established in table 1. Any two conditions that are underlined by the same line may be considered statistically identical given a p threshold of 0.05. For instance, the line connecting S and T in Q3 of the conversation scene indicates that there were no significant differences in terms of overall believability between saliency model and tracked gaze.

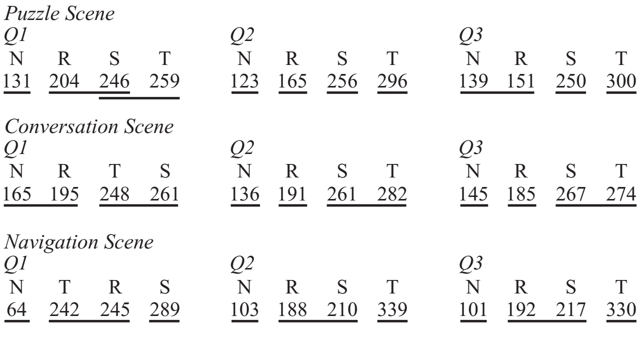


Figure 4: *Multiple comparison score for all data. Any conditions whose scores are underlined are considered statistically similar.*

The data captured was plotted in figure 5 for the tracked gaze, random model and saliency model. This enabled the comparison of the gaze conditions in terms of the correlation of the plots. The plots presented shows the frequencies of five gaze parameters (proximity, saccade magnitude, saccade velocity, fixation duration and the eccentricity) for one experienced participant. These plots serve to support the results from the evaluation experiment, given the limited data. Generally, the saliency model produced plots that were highly correlated to the tracked gaze while plots for the random model consistently underperforms.

6 Discussion

The hypothesis that the perceived realism of the avatar operating with the saliency model gaze would approach the ratings of the avatar exhibiting the actual tracked gaze was tested. It was expected that the avatar would be judged less favourably during the random model and static gaze conditions. Indeed, average rankings across scenes and questions indicate the superiority of tracked gaze, followed next by the saliency model, random and lastly static gaze. Overall, the plots (figure 5) for two parameters, saccade magnitude and velocities across all scenes produced were highly-correlated (i.e. correlation coefficient, $\rho_{T,S}$ ranged between 0.987 to 0.999) for the tracked gaze and the saliency model, as compared to the random model's less-correlated plots. While this overall picture indicates support for the saliency model and original hypothesis, the evaluation results and the other three parameters (proximity, fixation duration and eccentricity) must be discussed in terms of each scenario for a thorough examination.

Given that the saliency model was trained on data collected from the cubes puzzle scenario, the saliency model was rated significantly lower than tracked gaze in terms of eye and overall realism, but identical in terms of engagement. Random model was rated significantly lower than the saliency model in terms of realism, and the difference between these two conditions is greater than between tracked and saliency model by 211.1%. Similarly in terms of engagement, there was no significant difference tracked gaze and saliency model p=0.87. The plot of the eccentricity in this scenario shows a peak at the same point for the tracked gaze and the saliency model as compared to the random model. The proximity plot on the other hand, peaked at different points, although it's clearly more similar than the random model which slightly peaked at two points. Possible factors that can affect proximity include user habits such as how the user positioning when interacting with the cubes in the scenes. However, the plots do show that the saliency model was more likely to pick targets with the closest eccentricity and proximity to the tracked gaze. Clearly, more data is needed to make more sense of the fixation duration plot in this scenario.

The performance of the saliency model was statistically identical to ratings gained by actual tracked gaze during the conversational scene, as illustrated in figure 4. This affirms the success of the saliency model's approach to scene analysis towards realistic gaze generation. In this scene, random and static gaze are seen to significantly underperform. The peakedness of the eccentricity plot also demonstrates the viability of the target selection method employed by the saliency model. However, the proximity data also demonstrates the spatial clustering of the objects within the room. It must be noted that saliency model performed rather well in this scenario, possibly as a result of intense interaction during the conversation. Thus, the saliency model makes good use of more intrinsic saliency parameters such as the change in orientation and velocity of the other avatar in the room. This result is promising, in that the saliency model is more likely to perform well in highly interactive scenarios.

Finally, results from the navigation scene revealed that the saliency model was rated significantly higher than all other conditions when rating engagement. It is likely that ratings of engagement were likely to be informed by alternative stimuli, such as level of gaze activity relative to background activity within the scene. Percep-

Table 2: *Computed preference matrix for:*

(a) *object-focussed puzzle* scene.

	Q#	N	R	S	T	Total
N	1	-	53	43	35	131
	2	-	44	48	31	123
	3	-	58	51	30	139
R	1	87	-	58	59	204
	2	96	-	30	39	165
	3	82	-	35	34	151
S	1	97	82	-	67	246
	2	92	110	-	54	256
	3	89	105	-	56	250
T	1	105	81	73	-	259
	2	109	101	86	-	296
	3	110	106	84	-	300

(b) *conversation* scene.

	Q#	N	R	S	T	Total
N	1	-	63	57	45	165
	2	-	58	46	32	136
	3	-	59	47	39	145
R	1	82	-	47	66	195
	2	87	-	46	58	191
	3	86	-	44	55	185
S	1	88	98	-	75	261
	2	99	99	-	63	261
	3	99	101	-	67	267
T	1	99	79	70	-	248
	2	112	87	83	-	282
	3	105	90	79	-	274

(c) *navigation* scene.

	Q#	N	R	S	T	Total
N	1	-	24	21	19	64
	2	-	41	39	23	103
	3	-	39	36	26	101
R	1	116	-	54	75	245
	2	99	-	60	29	188
	3	101	-	57	34	192
S	1	119	86	-	84	289
	2	101	80	-	29	210
	3	104	83	-	30	217
T	1	121	65	56	-	242
	2	117	111	111	-	339
	3	114	106	110	-	330

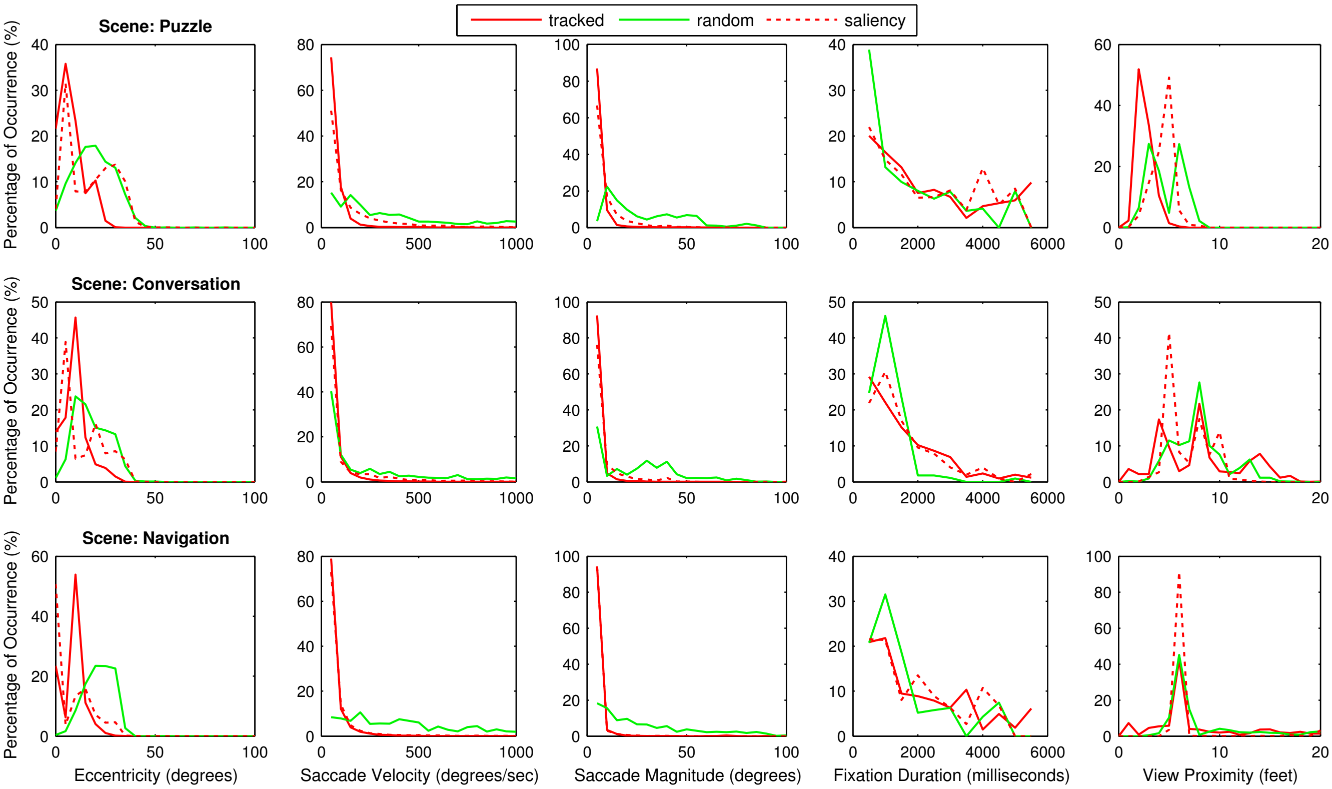


Figure 5: *Comparisons of gaze parmeters between scenes.*

tions of engagement may have been based on this factor. However, it must be noted that tracked gaze was rated significantly higher in terms of eye and overall realism. Despite the larger scale of the town scene and the wider spatial clustering of objects, the plots do show that the saliency model was more likely to pick targets with closer eccentricity and proximity to the tracked gaze. The peaks in the fixation duration plots at the 500-1000ms mark for the random model was clearly reduced for the conversation and navigation scenes demonstrating the viability of the saliency model's duration on target chosen.

7 Conclusions

Overall, the saliency model has yielded remarkably good performance on a wide variety of virtual reality scenes. This result was achieved by modelling the interactions between objects, rather than attempting to develop a model for specific environments and tasks. Further ongoing testing of the algorithm includes comparison between the intrinsic saliency criteria. The intrinsic saliency of the objects in the virtual scenes tend to interact and more research is needed into how to bias the relative weights of saliency parameters to tune systems towards specific virtual scenes. Itti et al [Itti et al. 1998] builds on their framework for interpreting complex natural scenes and suggest supervised learning as a strategy to bias the relative weights of the features in order to tune the system towards specific target detection tasks. Competing saliency effects in the virtual scene depends on the spatial characteristics of the target scene, hence the question of which saliency effects and when they should be implemented requires further research. Furthermore, the evaluation and the plots in figure 5 were not the saliency model's best measure of target relevancy. Future work will concentrate on designing a task to test the ability of the saliency model in detecting accurate targets within a virtual scene. This should also address a limitation of the saliency model's evaluation, in that it was generated from a single person's head gaze behaviour.

There are numerous extensions to this work that are worthy of further research. An interesting aspect of the work is to extend the model to allocate attention to surfaces of the objects instead of the center of the objects, as is currently implemented. Research is needed into how attention is allocated during object scrutiny. Secondly, the addition of a realistic head movement model is an obvious extension to the saliency model. Indeed, it may be extended to fully animate an avatar in its entirety i.e. pointing and locomotion. Obviously, the issue of where an avatar should point or move to is a separate problem. Thirdly, further studies can also influence designers on placement of objects within virtual scenes. Cues can also be used as phantom targets in virtual reality applications, movies and scenes. Another interesting extension to this work is to combine this saliency approach with cognitive or interaction-based approaches.

While this saliency modelling approach is not dependent on cognitive operations, it has the virtue of a straightforward implementation that can be applied to any virtual scene composed of a scene database of objects.

Acknowledgements

The authors acknowledge the support of EPSRC *Eye Catching* (EP/E007406/1) and EU *Presenccia* (contract no. 27731) Projects.

References

ARGYLE, M., AND COOK, M. 1976. *Gaze and Mutual Gaze*. Cambridge University Press Cambridge.

BENTIVOGLIO, A., BRESSMAN, S., CASSETTA, E., CARRETTA, D., TONALI, P., AND ALBANESE, A. 1997. Analysis of blink rate patterns in normal subjects. *Movement Disorders 12*, 6.

FINDLAY, J., AND WALKER, R. 1999. A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences 22*, 04, 661–674.

GARAU, M., SLATER, M., VINAYAGAMOORTHY, V., BROGNI, A., STEED, A., AND SASSE, M. 2003. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Human factors in computing systems*, 529–536.

GRILLON, H., AND THALMANN, D. 2009. Simulating gaze attention behaviors for crowds. *Computer Animation and Virtual Worlds, 20 2*, 3, 111–119.

GU, E., AND BADLER, N. 2006. Visual attention and eye gaze during multiparty conversations with distractions. *Lecture Notes in Computer Science 4133*, 193.

HENDERSON, J., AND HOLLINGWORTH, A. 1999. High-level scene perception. *Annual Review of Psychology 50*, 1, 243–271.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence 20*, 11, 1254–1259.

ITTI, L., DHAVALE, N., AND PIGHIN, F. 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. of the SPIE 48th Annual International Symposium on Optical Science and Technology*, 64–78.

JAMES, W. 1890. *The principles of psychology*.

KENDALL, M., AND SMITH, B. 1940. On the method of paired comparisons. *Biometrika 31*, 3-4, 324–345.

KHULLAR, S., AND BADLER, N. 2001. Where to Look? Automating Attending Behaviors of Virtual Human Characters. *Autonomous Agents and Multi-Agent Systems 4*, 1, 9–23.

KOCH, C., AND ULLMAN, S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology 4*, 4, 219–227.

LEDDA, P., CHALMERS, A., TROSCIANKO, T., AND SEETZEN, H. 2005. Evaluation of tone mapping operators using a high dynamic range display. *Proceedings of ACM SIGGRAPH 2005 24*, 3, 640–648.

LEE, S., BADLER, J., AND BADLER, N. 2002. Eyes alive. *ACM Transactions on Graphics (TOG) 21*, 3, 637–644.

LEE, J., MARSELLA, S., TRAUM, D., GRATCH, J., AND LANCE, B. 2007. The rickel gaze model: A window on the mind of a virtual human. *Lecture Notes in Computer Science 4722*, 296.

LUEBKE, D. 2003. *Level of detail for 3D graphics*. Morgan Kaufmann.

MA, X., AND DENG, Z. 2009. Natural Eye Motion Synthesis by Modeling Gaze-Head Coupling. In *Proc. of IEEE Virtual Reality Conference*, 143–150.

MASUKO, S., AND HOSHINO, J. 2007. Head-eye Animation Corresponding to a Conversation for CG Characters. In *Computer Graphics Forum*, vol. 26, Blackwell Synergy, 303–312.

MELCHER, D., AND KOWLER, E. 2001. Visual scene memory and the guidance of saccadic eye movements. *Vision Research 41*, 25-26, 3597–3611.

MURGIA, A., WOLFF, R., STEPTOE, W., SHARKEY, P., ROBERTS, D., GUIMARAES, E., STEED, A., AND RAE, J. 2008. A Tool For Replay And Analysis of Gaze-Enhanced Multiparty Sessions Captured in Immersive Collaborative Environments. In *Proc. IEEE/ACM DS-RT 2008*, 252–258.

PETERS, C., PELACHAUD, C., BEVACQUA, E., MANCINI, M., AND POGGI, I. 2005. A model of attention and interest using gaze behavior. *Lecture notes in computer science 3661*, 229.

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

STENTIFORD, F. 2007. Attention-based similarity. *Pattern Recognition 40*, 3, 771–783.

STEPTOE, W., OYEKOYA, O., MURGIA, A., WOLFF, R., RAE, J., GUIMARAES, E., ROBERTS, D., AND STEED, A. 2009. Eye Tracking for Avatar Eye Gaze Control During Object-Focused Multiparty Interaction in Immersive Collaborative Virtual Environments. *IEEE Virtual Reality Conference, 2009.*, 83–90.

TREISMAN, A., AND GELADE, G. 1980. A feature-integration theory of attention. *Cognitive psychology 12*, 1, 97–136.

VINAYAGAMOORTHY, V., GARAU, M., STEED, A., AND SLATER, M. 2004. An Eye Gaze Model for Dyadic Interaction in an Immersive Virtual Environment: Practice and Experience. *Computer Graphics Forum 23*, 1, 1–11.

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

YARBUS, A. 1967. *Eye movements and vision*. Plenum press.