EUROGRAPHICS 2002 / I. Navazo Alvaro and Ph. Slusallek
(Guest Editors)

Modeling Visual Attention in VR: Measuring the Accuracy of Predicted Scanpaths

Gerd Marmitt† and Andrew T. Duchowski†

Department of Computer Science, Clemson University, Clemson, SC, USA

Abstract

Dynamic human vision is an important contributing factor to the design of perceptually-based Virtual Reality. A common strategy relies on either an implicit assumption or explicit measurement of gaze direction. Given the spatial location of foveal vision, computational resources are directed at enhancing the foveated region in real-time. To obtain an explicit gaze measurement, an eye tracker may be used. In the absence of an eye tracker, a computational model of visual attention may be substituted to predict visually salient features. The fidelity of the resultant real-time system hinges on the agreement between predicted and actual regions foveated by the human. The contributions of this paper are the development and evaluation of a novel method for the comparison of human and artificial scanpaths recorded in VR. The novelty of the present approach is the application of previous accuracy measures to scanpath comparison in VR where analysis is complicated by head movements and dynamic imagery. An attentional model previously used for view-dependent enhancement of Virtual Reality is evaluated. Analysis shows that the correlation between human and artificial scanpaths is much lower than expected. Recommendations are made for improvements to the model to foster closer correspondence to human attentional patterns in VR.

1. Introduction

To increase the visual fidelity of Virtual Environments (VEs), recent strategies have emerged which combine view-independent rendering solutions (e.g., global illumination) with view-dependent perceptually-driven enhancements. Perceptually-driven techniques generally exploit the perceptual limitations of the Human Visual System (HVS) and reduce the computational burden by either withholding imperceptible information (e.g., compression) or by focusing computational resources within highly salient regions. In Virtual Reality (VR), the latter form of perceptually-driven enhancement typically relies on either an implicit assumption or explicit measurement of the human's instantaneous foveal Region Of Interest (ROI) or Point Of Regard (POR). Given the spatial location of the human POR, computational resources are directed at enhancing the foveated region in real-time by, for example, ray tracing a region of limited spatial extent. This provides a "just-in-time" view-dependent enhancement of the pre-rendered view-independent scene.

The key to "just-in-time" view-dependent enhancement is the determination of the instantaneous ROI to which the human participant is attending. To obtain an explicit measurement of the human's foveal ROI, an eye tracker may be used. In the absence of an eye tracker, a computational model of visual attention may be substituted to predict instantaneous human Regions Of Interest. The fidelity of the resultant real-time system hinges on the agreement between predicted and actual eye movements made by the human.

The aim of this paper is to measure the accuracy of an attentional model previously used in Virtual Reality for the purpose of real-time view-dependent scene enhancement. The contributions of this paper are the development and evaluation of a novel method for the comparison of human and artificial scanpaths following immersion in a VE. The comparison method is based on the well-known scanpath similarity indices derived from *string editing*. Scanpath similarity indices have successfully been used to measure the accuracy of visual attention models over still images. The novelty of the present approach is the application of the accuracy measures to human scanpaths in VR where the comparison is complicated by head movements and dynamic imagery.

† {gmarmit | andrewd}@vr.clemson.edu

© The Eurographics Association 2002.

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

2. Background

Perceptually-based rendering incurs a tradeoff between minimizing computational resources and maximizing (or preserving) image quality as perceived by a human observer. Many approaches have been considered. For real-time applications such as Virtual Reality, a particularly relevant class of perceptually-based rendering algorithms involves approaches where the instantaneous location of a participant's gaze is used to guide high-fidelity rendering approaches in a *gaze-contingent* manner. In these applications, an eye tracker is often used to determine the participant's direction of gaze.[7, 8] If an eye tracker is unavailable, a computational model of visual attention may be substituted. The modeling approach is a popular alternative in static image rendering and has recently been applied to real-time rendering of a Virtual Environment.[3] Among several decision criteria, Haber et al. employed a model of visual attention patterned after the model developed by Itti et al.[5] to predict the most salient objects within a scene assumed to attract the participant's visual attention. Based on these artificially determined Regions Of Interest (aROIs) the globally illuminated scene was partially updated with a real-time ray tracer restricted to shooting rays within the small (foveal) aROI.

Haber et al.'s perceptually-guided corrective splatting algorithm is one of the first VR applications to employ a real-time model of visual attention. The model operates by evaluating a projected image on a frame-by-frame basis. While Itti et al.'s model has been shown to be quite accurate over still images, it is not clear how well the model generalizes to the dynamic scene content presented during VR immersion.

2.1. Visual Attention Modeling

Building on biologically-plausible architectures of the human visual system, the model developed by Itti et al.[5] is related to Treisman's Feature Integration Theory.[12] Starting with an input image, it is progressively low-pass filtered and subsampled to yield nine dyadic spatial scales. The multi-scale image representation is then decomposed into a set of topographic feature maps. Each feature is computed via a set of linear "center-surround" operations akin to visual receptive fields. Different spatial locations then compete for saliency within each map, such that only locations which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master "saliency map". The saliency map contains internal dynamics which generate attentional shifts. The model has been tested on a variety of artificial and natural images and appears to be very robust, particularly to the addition of noise. The model's performance is in general consistent with observations in humans and is consistent with Treisman's Feature Integration Theory.

For adaptation of Itti et al.'s model to real-time VR rendering, all steps up to the computation of the saliency map need only to be performed once per image. This map is used for computing all aROIs in the image. Computing the saliency map is the most expensive operation. In general, without any modification, the algorithm does not necessarily support real-time applications. To reduce computational time, several components of the model may be removed to guarantee real-time operation. For example, orientation map processing may be omitted, as suggested by Haber et al.[3]

2.2. Comparison of Human and Artificial ROIs

To test any model of visual attention, one very attractive methodology is to compare the sequence of Regions Of Interest (ROIs) identified by an attentional algorithm to those foveated by human observers. A recent study by Privitera and Stark presents just such a methodology for comparing algorithmic ROIs, or aROIs to those selected by humans, hROIs.[10] The comparison algorithm relies on two important processes: one of clustering of ROIs for comparison of loci of ROIs and a subsequent step of assembling the temporal sequences of ROIs into ordered strings of characters for comparison of sequences based on *string editing*. String editing, defined by an optimization algorithm based on the Levenshtein distance,[11] assigns unit cost to three different character operations: *deletion*, *insertion*, and *substitution*. Characters are manipulated to transform one string to another, and character manipulation costs are tabulated to yield a sequence similarity index S_s. A positional, or loci, similarity index S_p can be found for two strings by examining the characters of the second string to those of the first. Similarity coefficients are sorted and stored in a table, named the Y-matrix, having as many rows and columns as the number of different sequence ROIs to be considered.

Scanpath comparison values from the Y-matrix (which typically contains large amounts of data) are condensed (averaged) and reported in two tables, called Parsing Diagrams, one for each of S_s and S_p indices. Each of the parsing diagrams reports several correlation measures: idiosyncratic, local, and global. Idiosyncratic values report on the within-subject attentional scanning tendencies of individual subjects, i.e., these values report correlation measures between scanpaths made over different pictures by the same subject. For example, in reading studies, English readers would be expected to exhibit high idiosyncratic indices due to the adopted left-to-right text scanning strategies. Local indices report on between-subject correlations of scanning patterns over similar stimuli. That is, local indices report on different subjects' scanpaths over the same picture. In the present case, these are the most important results since these indices divulge the correlation between human and artificial scanpaths made over the same images. Global measures report on the correlation between scanpaths made by different subjects over different stimuli. Should these values be highly correlated (for hROIs made by different people or aROIs and hROIs over different environments), this would suggest that

© The Eurographics Association 2002.

S_p	Subj. 1		Subj. 2	
	Pict1	Pict2	Pict 1	Pict 2
S1P1	**R**	**I**	**L**	**G**
S1P2		**R**	**G**	**L**
S2P1			**R**	**I**
S2P2				**R**

	Same Subj.	Diff. Subj.
Same Image (SI)	**R**epetitive	**L**ocal
Diff. Image (DI)	**I**diosyncratic	**G**lobal
S_p	Random	

S_s	Subj. 1		Subj. 2	
	Pict1	Pict2	Pict 1	Pict 2
S1P1	**R**	**I**	**L**	**G**
S1P2		**R**	**G**	**L**
S2P1			**R**	**I**
S2P2				**R**

	Same Subj.	Diff. Subj.
	Repetitive	**L**ocal
	Idiosyncratic	**G**lobal
S_s	Random	

Figure 1: *Adaptations of Y-matrices and parsing diagrams.*[10]

Virtual Environments tend to be viewed similarly by different people (and by the attentional model). Examples of a Y-matrix and parsing diagrams are shown in Figure 1.

Using these similarity measures, Privitera and Stark evaluated ten different attentional algorithms. In general, although the set of tested algorithms was only a small representative sample of many possible procedures, this set could indeed predict eye fixations over still images. It appears that a multiresolutional strategy, such as that of Itti et al., seems to be very efficient for several classes of images.

3. Methodology for Model Evaluation in VR

The string editing approach to scanpath comparison is defined to operate over two scanpaths captured over still images. To permit comparison of scanpaths generated while immersed in VR, the methodology is applied to sets of still images obtained during immersion. Thus to allow use of the scanpath comparison technique, a participant's immersive session in VR must be analyzed to locate periods of time when the image viewed by the participant is relatively still.

Two novel methods of temporal analysis are presented: *head-based* and *time-based*. Head-based analysis is used to locate sequences of still images where the head is stable (but the eyes may not be). Based on estimates of head stability, this analysis technique presents a still image to the attentional model for variable periods of viewing time. To more closely resemble the real-time use of the attentional model in VR, the time-based analysis approach assumes a constant frame rate (10 fps). In this approach, the attentional model is modified in a manner similar to Haber et al. so that it is constrained to locating aROIs within a constant time period (100 ms). The difference between the two approaches lies in the amount of time allotted to the attentional model to locate aROIs. The head-based approach favors the attentional model (since the model is given more time to analyze an image) while the time-based technique better mimics real-time constraints placed on the attentional model in a VR application.

To allow comparison between aROIs and hROIs in both approaches, human fixations are identified via velocity-based analysis of eye movements over identical sequences of images as those input to the attentional model.

3.1. Head-Based Analysis

The head-based analysis methodology aims at isolating periods of immersion in VE where the head (and hence image) is stable. The resulting sequence of image frames, averaged to a single image, is used to collect human and artificial eye movements over the period of (relative) head stability. Resulting scanpaths are then evaluated against automatically located ROIs by Itti et al.'s attentional model over variable viewing periods.

As suggested by Jaekl et al.,[6] to analyze head motion, head movements are considered separately within 3 degrees of freedom in terms of their translational and rotational movement components. To establish appropriate thresholds and to examine captured signal components, three benchmark head movements were recorded: stable, slow, and fast rotation. The captured head rotation signals are shown in Figure 2. The angular velocity plots have been processed to remove signal noise.

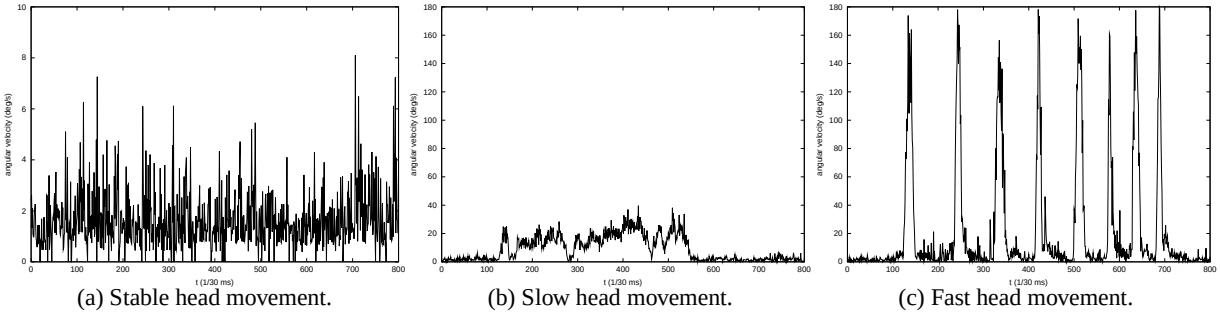
The rotational head-stable detection algorithm calculates the degree of change for Euler angles roll (φ), elevation (or pitch, ρ), and azimuth (or yaw, γ). To obtain angular velocity (ω, in deg/s), the difference between two successive orientations is calculated as a 3-vector and divided by the time between samples. That is, given the instantaneous orientation vector $\phi_i = (\varphi_i, \rho_i, \gamma_i)$,

$$\omega = \frac{\| \phi_{i+1} - \phi_i \|}{\Delta t} \quad \text{deg/s}$$

where Δt is the time between samples. A velocity threshold of 23 deg/s was chosen to locate fast changes in head rotation. Data between these fast movements were then considered as sequences of rotational head stability.

To complete the head-stable analysis, a translational movement threshold is needed. A velocity thresholding filtering approach was chosen, where velocity is determined as the difference between two spatial head positions divided by the time between samples. Given two successive

© The Eurographics Association 2002.



(a) Stable head movement. (b) Slow head movement. (c) Fast head movement.

Figure 2: *Angular head velocity (in deg/s).*

head positions in three-space, $\mathbf{p}_i = (x_i, y_i, z_i)$ and $\mathbf{p}_{i+1} = (x_{i+1}, y_{i+1}, z_{i+1})$, velocity is calculated as

$$\nu = \frac{\| \mathbf{p}_{i+1} - \mathbf{p}_i \|}{\Delta t} \quad \text{ft/s}$$

where Δt is the time between samples. A 3 ft/s (91.44 cm/s) threshold was selected empirically. It should be noted that the translational threshold is only used to support the localization of head-stable sequences—for the purposes of determining head-stable imagery, the more important consideration is the head's angular velocity.

In cases where subjects performed slow head rotations below threshold, head-stable sequences were found by splitting the entire sequence based on detected absolute changes of azimuth and elevation exceeding 5 deg/s or a detected roll exceeding 2.5 deg/s. To summarize, the following four conditions define stable head movement: (1) $\omega \leq T_\omega$, (2) $\nu \leq T_\nu$, (3) $\mid \rho + \gamma \mid < A$, (4) $\mid \varphi \mid < B$, with the following parameters: T_ω = 23 deg/s, T_ν = 3 ft/s (91.44 cm/s), A = 5 deg/s, and B = 2.5 deg/s.

3.2. Time-Based Analysis

In contrast to the above head-based analysis, where viewing periods are variable, a time-based analysis approach was developed to evaluate the accuracy of Itti et al.'s model in a real-time setting where images of constant duration (based on frame rate) are presented to the viewer. To match the real-time frame rate of 10-12 fps reported by Haber et al., the performance of Itti et al.'s algorithm was examined. Following Haber et al.'s recommendations, tests showed that Itti et al.'s algorithm can be made in most cases to extract the most salient region within 100 ms. While a typical extent of 5° visual angle is associated with foveal vision, it is not clear how close human eye movements (or human ROIs) fall to the automatically located attentional regions (aROIs). To evaluate the average distance between aROIs and hROIs, the same raw data as used in the head-stable analysis (collected from immersive trials; see below) is analyzed to evaluate the attentional model's accuracy. This time, instead of identifying sequences of images displayed during stable head movements, hROIs and aROIs are compared over frames collected every 100 ms.

In this form of analysis, the only processing step required is to identify and regenerate those image frames that were shown to participants during active use of the eye tracker (i.e., frames shown to the participant when the eye tracker is either in reset of calibration mode are not considered).

To enable comparison of aROIs and hROIs over sequences of images shown during 100 ms cycles, eye movement data is analyzed over image frames to locate fixations (see below). If, on a given frame, no fixation is detected (i.e., the subject was performing a saccade), no comparison is possible. That is, image frames where no human or artificial fixations are detected are not considered in the comparative analysis.

Note that the simplifications implemented in the time-based analysis are aimed at simulating real-time use of an attentional model (e.g., in VR). In this case, due to time constraints and expected lengths of scanpaths (usually 1 fixation), building string sequences for comparison makes no sense if there are only two fixations per frame to compare. Instead of scanpath sequences, the time-based analysis relies on the computation of pixel distance between aROIs and hROIs in the image. The idea here is to provide an approximation of the extent of the area required for update by a ray tracer or other similar perceptually-based enhancement. It is expected that such a real-time foveal window should be limited in size but still offer the same perceptual impression as that of a fully ray traced environment displayed at the same frame rate (which is currently not possible and is of course the reason for gaze-contingent systems). In this analysis, instead of employing scanpath comparison measures used by Privitera and Stark, or reporting a single distance measure per frame, time-based analysis produces a table reporting the size of the model's attentional window needed to overlap human fixations over a set percentage of frames.

3.3. Eye Movement Analysis

In both head- and time-based analysis approaches, an intrinsic component of scanpath comparisons relies on the analysis of eye movements. Due to saccadic suppression, generally the most meaningful eye movements are those where the viewer is fixating steadily during inspection of

© The Eurographics Association 2002.

the scene. Eye movement analysis, therefore, is concerned with the identification of fixations. To locate fixations, an acceleration-based velocity filtering approach with adaptive thresholding is chosen to detect saccades. Fixations are then classified as those periods in the eye movement data stream that occur between saccades. A modified version of a 3D eye movement algorithm is used.[2] Because only projected 2D images of the 3D environment are usable by the attentional model, only monocular 2D eye movements made on the near focal plane are considered for analysis.

3.4. Scanpath Comparison

The segmentation of image frames by either head- or time-based methods provides image sequences over which human and artificial scanpaths can be compared. In head-based analysis, following identification and grouping of stable head positions and orientations, identified fixations are assigned to the appropriate head-stable sequence of images (i.e., images when the head was stable and when fixations were detected). If a given eye fixation spans two (or more) head-stable image groups, the fixation is assigned to both (or all) groups. Prior to this duplication of overlapping fixations, most image sequences contained only one or two fixations (which was not surprising due to the relatively short durations of identified head stability). To obtain eye movement results from the attentional model, averages of head position, up, and view vectors are used to generate a single image used as input to the attentional algorithm. To take into account the possibility of the Vestibulo-Ocular Response (VOR, a type of eye movement used to stabilize a retinal image during head motion), an average of the viewer's gaze is obtained by bisecting the gaze direction at the beginning and end of the head-stable image sequence. This bisected vector forms the direction of view from which an input image is generated. The attentional model is limited to run in the same time as the duration of the head-stable sequence.

The resulting human and artificial scanpaths over sequences of images are then compared in a manner similar to the technique proposed by Privitera and Stark,[10] with two key differences. First, repetitive measures (same subject, same pictures) are not calculated since, even in the case of several repeated subject trials, identical images (based on head position and orientation) can not readily be reproduced. To do so would require each subject to exactly repeat their previous navigational sequence within the environment. Second, to facilitate string editing, instead of a *k*-means clustering approach to fixation grouping used by Privitera and Stark, fixation grouping is performed by defining a 5° spatial window for clustering. If two or more human (or algorithmic) fixations fall within this window, fixations are clustered and relabeled as a new fixation located at the centroid of the cluster. This within-sequence clustering procedure effectively reduces a large number of fixations (in either human or algorithmic sequence). A similar spatial clustering strategy is employed to label fixations for subsequent between-sequence string editing comparisons. That is, if the attentional model generates fixations within 5° distance to the center of a human fixation, the algorithmic and human fixations are counted as falling within the same fixation group (a measure of equivalence, resulting in the same character label for string editing). All other unassigned algorithmic fixations are assigned new character labels.

Further scanpath analysis is identical to Privitera and Stark's. Loci (S_p) and path (S_s) similarity indices are calculated for idiosyncratic (same subject, different image), local (different subject, same image), and global (different subject, different image) scanpath comparisons.

4. Results

Attempting to maximize the model's attentional performance at lowest computational cost, four variants of the attentional model are compared to human scanpaths. All variations of the algorithm rely on a key modification which alters the amount of time allowed to process image sequences. In some cases, the attentional model is given certain advantages over the human, e.g., extra time to compute artificial Regions Of Interest. All but one variant of the model generate artificial scanpaths limited to the number of fixations detected by the human over a given frame. The four model variations are described as follows:

1. Case 1: Pseudo-Time Measurement. In this case, the algorithm is allowed to use its own time measurement to finish processing in the time allotted (e.g., in head-based analysis, time is based on periods of head stability). Note that the algorithmic time measurement is only a pseudo-time measurement, since it mainly represents the number of iterative steps made by the internal neural network rather than the real-time amount. In this variant, this measurement excludes the calculation of the saliency map.
2. Case 2: Identical Number of Fixations. In head-based analysis, the attentional model is given as much time as required, limited to the number of fixations made by the human.
3. Case 3: Near-Real-Time. In this case, instead of allowing the attentional model to use its own intrinsic timing facility, the model is timed using the independent system function `gettimeofday()`. Note that this measurement is not quite real-time since the measurement is made on a multi-tasking Unix platform. To ensure near-real-time performance in this (and the next case), the model's neural network is replaced with a mechanism to locate the most salient pixel region in the saliency map (i.e., maximum luminance detection). Because this modification significantly enhances the model's performance, the model is capable of generating hundreds of fixations in near-real-time. This is clearly an oversimplification of the model since localization of this many fixations defeats the purpose of attentional analysis of the image (with this

© The Eurographics Association 2002.

many fixations, one may as well render the entire scene at full fidelity). For this reason, since it is assumed there is an underling attentional mechanism represented by the saliency map, the model is restricted to selecting as many fixations as detected by the human. Furthermore, in this case, the time taken by the model to generate the saliency map is not taken into consideration.

4. Case 4: Full-Real-Time (Picture rescaled to 1/4). In this case, the algorithm was examined for possible modifications leading to real-time performance. One of the most time-consuming components of the algorithm is its generation of the saliency map. This step may take normally over one second (on the available Unix platform). Omitting the computation of the orientation map (as suggested by Haber et al.[3]), execution time still averaged around 700 ms. Further time savings could only be made by rescaling the image and interpolating the result to its original size following algorithmic analysis. In this case the image is subsampled to 1/4 its original size (maintaining aspect ratio).

In all comparisons of model variations, human data is identical since the various approaches do not alter human performance. Random scanpath values are also not altered. The experimental conditions used to generate raw eye movement data were the same for both head-based and time-based analysis.

4.1. Experimental Design

A 3×3 factorial design was used, with three factors (three types of environments; cube, panoramic, CG; see below) at three levels, distributed between three trial groups of three subjects. Each group was assigned each type of environment, with each type of environment limited to viewing by only one group. Order effects were counterbalanced by organizing each group's viewing order in a 3×3 Latin square.

Apparatus. All experimental trials were conducted in the VR Lab at Clemson University. The primary rendering engine is a dual-rack, dual-pipe, Silicon Graphics Onyx2® InfiniteReality2™ system with 8 raster managers and 8 MIPS® R12000™ processors, each with 8MB secondary cache.† It is equipped with 8Gb of main memory and 0.5Gb of texture memory.

Multi-modal hardware components include a binocular eye tracker mounted within a Virtual Research V8 Head Mounted Display. The V8 HMD offers 640×480 pixel resolution per eye with individual left and right eye feeds. HMD position and orientation tracking is provided by an Ascension 6 Degree-Of-Freedom (6DOF) Flock Of Birds (FOB). The HMD is shown in Figure 3(inset), with the FOB sensor just visible on top of the helmet.

† Silicon Graphics, Onyx2, InfiniteReality, are registered trademarks of Silicon Graphics, Inc.

The eye tracker is a video-based, corneal reflection unit, built jointly by Virtual Research and ISCAN. Each of the binocular video eye trackers is composed of a miniature camera and infrared light sources, with the dual optics assemblies connected to a dedicated personal computer (PC). The ISCAN RK-726PCI High Resolution Pupil/Corneal Reflection Processor uses corneal reflections (first Purkinje images) of infra-red LEDs mounted within the helmet to measure eye movements. Figure 3 shows the dual cameras and infra-red LEDs of the binocular assembly. Mounted below

Figure 3: *Binocular eye tracker optics (w/HMD inset).*

the HMD lenses, the eye imaging cameras peer upwards through a hole cut into the lens stem, capturing images of the eyes reflected by a dichroic mirror placed behind the HMD lenses. The processor typically operates at a sample rate of 60Hz, however while in binocular mode the measured sample rate decreases to 30Hz. The subject's eye position is determined with an accuracy of approximately 0.3 degrees over a ± 20 degree horizontal and vertical range using the pupil/corneal reflection difference. The maximum spatial resolution of the calculated Point Of Regard (POR) provided by the tracker is 512×512 pixels per eye.

Subjects. Nine subjects were invited to participate in the experiment (7 males, 2 females; average age 26). Most of the subjects (8) were undergraduate students with no experience in Virtual Environments. Each subject was asked to fill out an informed consent form (approved by Clemson's Institutional Review Board), and was asked to indicate which eye was dominant (by asking to simulate shooting a rifle).

Procedure. Following an introduction and a brief description of the experiment, participants were asked to don the HMD and were given a brief preview of the types of environments in a short training session. Subjects first viewed each of the three types of environments in the prescribed order, but for eye movement data collection, the three training environments were chosen from a different group. Of particular interest are fixations during a first-time immersion in a VE. To avoid memory effects, each subject was exposed to a new environment, and each environment was viewed only once. Exposure to environments was limited to the following: Level 0: 10 sec; Level 1: 20 sec; Level 2: 40 sec. Durations for viewing training environments were not

© The Eurographics Association 2002.

as strict. Each subject viewed the training environments until they were comfortable in each.

Before viewing each test environment, each participant underwent a 5-point eye tracker calibration sequence. Calibration quality of each eye was noted (to facilitate subsequent eye movement analysis). In most cases, calibration was better for the subject's dominant eye (self-reported). Each trial lasted approximately 15 to 25 minutes. Each participant was asked whether they wanted to take a break after seeing the training environments or to continue with the trials. Most subjects chose to continue immediately.

No particular task was assigned during trials, e.g., subjects were allowed to "free view" the environment. However, one problem was observed when viewing the CG scene environments: all CG environments lacked a wall, creating an open space at one end of the environment. If subjects were seen looking at this empty space, they were asked to divert their attention toward more interesting parts of the environment (e.g., they were informed that there was something else to look at and that they were free to move their head about). Anecdotal analysis indicates there at least two types of VR participants: some exhibit very slow head movements, usually not being able to view the entire environment in the time allotted, while others explore the environment with apparently great enthusiasm and cover the entire scene within a few seconds (generating fast head movements).

Stimuli. To gauge eye movements in VR and to mimic the environment generated by Haber et al., three types of VEs were used, shown in Figure 4, each of varying complexity:

- Level 0: A simple cube environment. The first type of environment is a simple box-like volume with simple stimuli texture-mapped on its walls. This environment, shown in Figure 4(a), was constructed from simple images over which the attentional model is known to perform well. All six inner faces contained the same texture. In all, three different textures were used (obtained from the publicly available distribution of Itti et al.'s code) in each of three environments.
- Level 1: A panoramic environment. To enable comparison of scanpaths over "natural" imagery, textures of natural scenes were used to create a 360° panoramic VE. These environments are similar to commonly seen QuickTime VR environments on the web. Since the environments were created by texture mapping a cylinder, the floor and ceiling of the environments were simply made of a homogeneous color (usually matching the dominant hue of the bottom and upper portions of the panoramic scenes, e.g., a blue ceiling for an "outdoor" environment). The 360° panoramic images were obtained from the web and mapped to a cylinder spanning the main vertical axis, so as to impart the impression of standing in the middle of the scene (see Figure 4(b)). All subjects were asked to minimize their head movements to rotations about the vertical axis (e.g., not to look down or up at the floor and ceiling). The three textures used represented a landscape, county fair, and fort.
- Level 2: A CG scene. To extend the complexity of scenes to purely synthetic scenes, and to simulate the work of Haber et al. (without perceptually adaptive enhancements), 3D radiosity environments were generated. All virtual environments were created at Clemson University using an in-house file format as well as in-house tools for computing radiosity-based global illumination. All VEs were fairly simple, consisting of one room and a few objects, e.g., table, light(s), chair(s). The most sophisticated of the three environments is shown in Figure 4(c).

4.2. Head-Based Analysis

Selected scanpath examples over three different VEs viewed in the experiment are shown in Figure 5. In these images, hROIs and aROIs are symbolized by circles and squares, respectively, each representing 5° visual angle coverage.

Average results are reported comparing human and artificial scanpaths in VR, following Privitera and Stark's string editing methodology. Results from comparisons of human scanpaths to those generated by each attentional modeling variant over each environment are reported in two parsing diagrams, given in Tables 1-4, one giving loci (S_p) correlation measures, the other giving scanpath order (S_s) similarity indices. The parsing diagrams are adapted to the present evaluation of two characteristic subjects: human and algorithm. For ease of comparison, human idiosyncratic indices are replicated in each of the four tables. Compared to the attentional model, human idiosyncratic indices are fairly highly correlated (at about 15%), roughly matching expected human idiosyncratic measures reported by Privitera and Stark. To facilitate testing for statistical significance between local indices, a random scanpath is generated to establish a "control" or baseline measurement composed of random fixations. The bottom-right entry of the parsing diagrams shows the local correlation between human and random fixations, i.e., correlation of scanpaths made by humans and a random process over the same image. Since viewing time for the random process is meaningless, the random process is simply made to generate the same number of fixations per image as the human.

In terms of human-artificial scanpath comparisons, the most salient values found in the parsing diagrams listed in Tables 1-4 are the local (different subject, same image) scanpath loci indices (S_p), listed as the upper-right element of each parsing diagram. Inspection of these values suggests that Cases 3 (Near-Real-Time) and 4 (Full-Real-Time) generate better agreement between aROIs and hROIs. This is somewhat surprising since those variants of the model which are given more time to analyze images were expected to provide better agreement. The reason for better performance of the Case 3 and 4 variants may be the removal of the orientation map, the same component removed by Haber et al. in

© The Eurographics Association 2002.

(a) A Level-0 scene (cube). (b) A Level-1 scene (panorama). (c) A Level-2 scene (CG environment).

Figure 4: *Example virtual environments.*

	Same Subj. (h)	Same Subj. (m)	Diff. Subj.	Same Subj. (h)	Same Subj. (m)	Diff. Subj.
L-0						
SI	×	×	**0.007**	×	×	**0.007**
DI	0.162	0.104	0.006	0.148	0.000	0.004
		S_p	0.044		S_s	0.000
L-1						
SI	×	×	**0.028**	×	×	**0.028**
DI	0.153	0.114	0.017	0.151	0.004	0.016
		S_p	0.031		S_s	0.013
L-2						
SI	×	×	**0.007**	×	×	**0.007**
DI	0.197	0.081	0.015	0.191	0.004	0.013
		S_p	0.020		S_s	0.017

Table 1: *Head-based analysis: Case 1*

	Same Subj. (h)	Same Subj. (m)	Diff. Subj.	Same Subj. (h)	Same Subj. (m)	Diff. Subj.
L-0						
SI	×	×	**0.003**	×	×	**0.003**
DI	0.162	0.015	0.022	0.148	0.001	0.015
		S_p	0.044		S_s	0.000
L-1						
SI	×	×	**0.015**	×	×	**0.015**
DI	0.153	0.005	0.016	0.151	0.003	0.015
		S_p	0.031		S_s	0.013
L-2						
SI	×	×	**0.023**	×	×	**0.021**
DI	0.197	0.019	0.026	0.191	0.010	0.021
		S_p	0.020		S_s	0.017

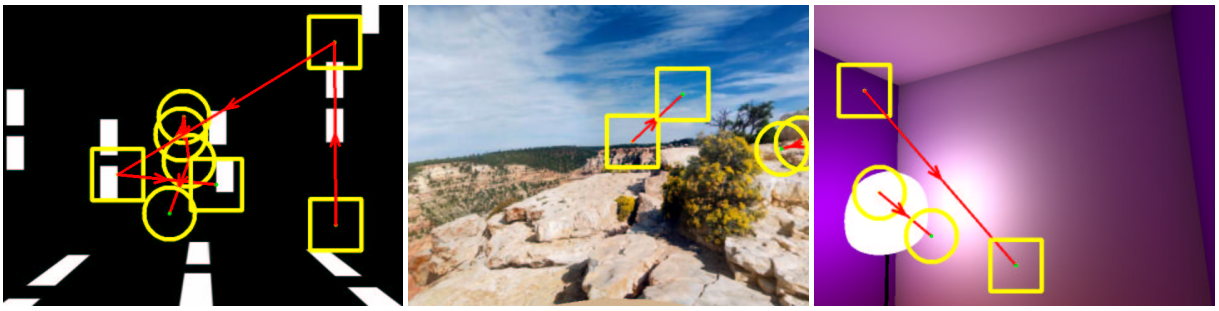
Table 2: *Head-based analysis: Case 2*

	Same Subj. (h)	Same Subj. (m)	Diff. Subj.	Same Subj. (h)	Same Subj. (m)	Diff. Subj.
L-0						
SI	×	×	**0.010**	×	×	**0.007**
DI	0.162	0.212	0.013	0.148	0.203	0.006
		S_p	0.044		S_s	0.000
L-1						
SI	×	×	**0.019**	×	×	**0.017**
DI	0.153	0.289	0.012	0.151	0.289	0.013
		S_p	0.031		S_s	0.009
L-2						
SI	×	×	**0.028**	×	×	**0.024**
DI	0.197	0.186	0.032	0.191	0.183	0.027
		S_p	0.020		S_s	0.017

Table 3: *Head-based analysis: Case 3*

	Same Subj. (h)	Same Subj. (m)	Diff. Subj.	Same Subj. (h)	Same Subj. (m)	Diff. Subj.
L-0						
SI	×	×	**0.026**	×	×	**0.026**
DI	0.162	0.100	0.018	0.148	0.086	0.010
		S_p	0.044		S_s	0.000
L-1						
SI	×	×	**0.020**	×	×	**0.020**
DI	0.153	0.060	0.016	0.151	0.058	0.015
		S_p	0.031		S_s	0.013
L-2						
SI	×	×	**0.016**	×	×	**0.014**
DI	0.197	0.110	0.022	0.191	0.102	0.017
		S_p	0.020		S_s	0.017

Table 4: *Head-based analysis: Case 4*



(a) Level-0 scene (cube). (b) Level-1 scene (panorama). (c) Level-2 scene (CG environment).

Figure 5: *Example scanpaths.*

© The Eurographics Association 2002.

	Case 1			Case 2			Case 3			Case 4			Random
	L-0	L-1	L-2	L-0	L-1	L-2	L-0	L-1	L-2	L-0	L-1	L-2	
100 %	48.30	42.70	43.38	46.69	42.09	43.17	41.70	41.73	49.42	54.76	47.49	48.20	48.26
90 %	43.10	36.75	37.31	41.01	36.43	37.03	35.06	35.87	44.34	48.82	41.43	43.92	42.03
80 %	36.99	32.70	32.45	36.44	32.34	32.48	32.32	32.86	41.70	44.41	37.10	41.26	37.37
70 %	34.23	29.24	28.81	33.51	28.47	28.69	28.66	28.97	38.42	39.10	32.95	38.34	32.95
60 %	30.97	25.75	25.42	30.42	25.37	25.18	26.34	25.78	35.20	34.11	29.35	35.50	29.72
50 %	28.08	22.61	21.57	26.09	22.25	21.83	22.34	22.52	33.36	30.05	26.02	31.47	25.98
40 %	23.76	19.39	18.24	22.01	19.20	18.25	19.20	19.84	31.06	26.18	22.17	28.74	22.33
30 %	18.23	16.09	14.65	18.57	15.67	14.21	15.86	16.58	28.93	20.82	19.23	25.59	17.87
20 %	14.31	11.38	10.18	13.35	11.19	10.03	11.29	13.05	26.60	16.69	14.90	21.48	14.34
10 %	4.01	2.86	2.03	3.56	3.45	1.87	5.77	3.55	12.67	5.64	5.28	12.53	3.74

Table 5: *Time-based analysis.*

their real-time implementation of a perceptually-based Virtual Environment.

4.3. Time-Based Analysis

As discussed in Section 3, the time-based analysis approach measures only the correlation between fixations since real-time frame display is used as the delimiting factor (100 ms per image). Here, instead of presentation of S_p and S_s values, results are given in terms of area coverage and distance between human and artificial fixations (in degrees visual angle). Table 5 lists average distance measures calculated for each environment under each modeling variant condition. The table should give an idea of how large the foveal region should be extended beyond the model's fixation point to match a specific amount (here percentage overlay) of human fixations. Lower numbers in the table indicate better performance (smaller distance between aROIs and hROIs). All time-based analysis techniques utilize the same variants of the attentional model as discussed above.

Examining average values reported in Table 5, perhaps the two most informative lines of data are at the 10% and 50% coverage levels. At the 10% level, relatively low values are seen for Cases 1 and 2 in comparison to those listed under Cases 3 and 4. Case 1 and 2 algorithm variants appear to perform slightly better than chance (compared to random ROI and hROI distances, reported in the far right column). Should the average distance values be significantly lower than random values at the 50% level, one might be tempted to claim that the attentional model is no different from another human (correlation between human subjects was reported at 50% by Privitera and Stark[10]). While statistical significance is not yet available, it does not appear likely that any variation of the model will qualify for this distinction.

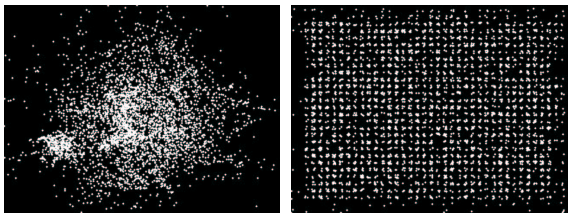
A curious trend can be seen in Table 5: unlike head-based analysis, at real-time constraints the model variants of Case 1 and Case 2 seem to outperform those of Case 3 and Case 4. This is again surprising since the opposite was expected. In real-time analysis, it may be that the lack of an orientation map at such minimal processing times (100 ms) may hinder algorithm performance. That is, since under real-time constraints generally only one fixation is expected per frame, it may be that the first ROI generated by Case 1 and 2 variants is generally more accurate than the first ROI generated by Case 3 and 4 variants. The orientation map may thus be more important at initial stages of the algorithm rather than at longer exposures.

5. Discussion

Analysis shows that the correlation between human and artificial scanpaths in all three types of environments is much lower than expected. At first, these results seemed rather incredulous. Certainly the attentional model appears to work quite well over still images, and as Privitera and Stark have shown, the techniques employed by the attentional model tend to identify foveal regions which generally agree with those identified by humans. However, it is the static nature of the stimulus that seems to be at issue. In most cases the image within head-stable or time-based sequences does not change much. The problem may lie in the algorithm's lack of memory. That is, each time the algorithm is run, as far as it is concerned, it is presented with a completely new image. In contrast, the human has already seen most parts of the image and is therefore free to distribute visual attention to new areas, even though, according to the algorithm's saliency map, the areas may appear to the model as less interesting.

Whatever viewing strategy is employed, calculated idiosyncratic indices suggest that humans tend to repeat adopted viewing patterns more consistently than the model. The attentional model appears to distribute attention to a wider area of the image. In contrast, humans appear to direct their attention to the central region of the image (at least, it is suspected, upon initial viewing or when viewing time is restricted). This is supported by distribution plots of human and artificial fixations, shown in Figure 6. Figure 6(a) shows that human fixations tend to cluster about the central image region, at least when immersed in VR. Indeed, this supports previous findings of human fixations being restricted to the central 30° of the VR display,[1, 13, 9] as well as findings of humans tending to fixate the image center within the first 2 or so seconds of initial image exposure.[14]

© The Eurographics Association 2002.



(a) Human fixations. (b) Artificial fixations.

Figure 6: *Spatial distribution of hROIs and aROIs.*

6. Conclusion and Recommendations for Future Work

The contributions of this paper are the development and evaluation of a novel method for the comparison of human and artificial scanpaths recorded in VR. String editing appears to offer a powerful analytical mechanism for evaluation of human-human or human-artificial scanpaths. An attentional model previously used for view-dependent enhancement of Virtual Reality is evaluated. Results indicate that, at least under immersive viewing conditions, the model does not accurately predict human fixations in VR. Despite the results presented here, it is believed that an image-based approach for real-time visual attention modeling in VR is not an unacceptable strategy. It is conjectured that any attentional model suitable for perceptually-based VR rendering should be augmented with a memory of previously seen image regions or perhaps previously inspected objects. This could be done by updating the saliency map dynamically over changing image regions, increasing the model's sensitivity to image transients while reducing its sensitivity to full-field transients (e.g., due to head motion).[4] The attentional model may be further enhanced by biasing it toward the central image location, at least during initial exposure.

Acknowledgments

This work was supported in part by NSF CAREER award # 9984278. We wish to thank Laurent Itti for providing us with the attentional model used in this study.

References

1. BARNES, G. R. Vestibulo-Ocular Function During Coordinated Head and Eye Movements to Acqurie Visual Targets. *Journal of Physiology* (1979), 127–147.

2. DUCHOWSKI, A., MEDLIN, E., COURNIA, N., GRAMOPADHYE, A., MELLOY, B., AND NAIR, S. 3D Eye Movement Analysis for VR Visual Inspection Training. In *Eye Tracking Research & Applications (ETRA)* (New Orleans, LA, March 25-27 2002), ACM, pp. 103–110,155.

3. HABER, J., MYSZKOWSKI, K., YAMAUCHI, H., AND SEIDEL, H.-P. Perceptually Guided Corrective Splatting. In *EuroGraphics* (Manchester, UK, September 4-7 2001), EuroGraphics.

4. ITTI, L. Real-Time High-Performance Attention Focusing in Outdoors Color Video Streams. In *Human Vision and Electronic Imaging VII (HVEI 02)* (San Jose, CA, 2002), SPIE, pp. 235–243.

5. ITTI, L., KOCH, C., AND NIEBUR, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 20*, 11 (1998), 1254–1259.

6. JAEKL, P. M., ALLISON, R. S., HARRIS, L. R., JASIOBEDZKA, U. T., JENKIN, H. L., JENKIN, M. R., ZACHER, J. E., AND ZIKOVITZ, D. C. Perceptual Stability During Head Movement in Virtual Reality. In *Virtual Reality (IEEE VR)* (Orlando, FL, March 24-28 2002), IEEE.

7. LUEBKE, D., HALLEN, B., NEWFIELD, D., AND WATSON, B. Perceptually Driven Simplification Using Gaze-Directed Rendering. Tech. Rep. CS-2000-04, University of Virginia, 2000.

8. MURPHY, H., AND DUCHOWSKI, A. T. Gaze-Contingent Level Of Detail. In *EuroGraphics (Short Presentations)* (Manchester, UK, September 4-7 2001), EuroGraphics.

9. MURPHY, H., AND DUCHOWSKI, A. T. Perceptual Gaze Extent & Level Of Detail in VR: Looking Outside the Box. In *Conference Abstracts and Applications (Sketches & Applications)* (San Antonio, TX, July 21-26 2002), ACM. Computer Graphics (SIGGRAPH) Annual Conference Series.

10. PRIVITERA, C. M., AND STARK, L. W. Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 22*, 9 (2000), 970–982.

11. SOCHER, G. `ftff, ftwhich and which-man: fault tolerant search utilities`. Guido's Linux Home-Page, URL: <http://main.linuxfocus.org/ guido/> (last accessed 04/06/02), 2001.

12. TREISMAN, A., AND GELADE, G. A Feature Integration Theory of Attention. *Cognitive Psychology 12* (1980), 97–136.

13. WATSON, B., WALKER, N., HODGES, L. F., AND WORDEN, A. Managing Level of Detail through Peripheral Degradation: Effects on Search Performance with a Head-Mounted Display. *ACM Transactions on Computer-Human Interaction 4*, 4 (December 1997), 323–346.

14. WOODING, D. S. Fixation Maps: Quantifying Eye-Movement Traces. In *Eye Tracking Research & Applications (ETRA)* (New Orleans, LA, March 25-27 2002), ACM, pp. 31–36.

© The Eurographics Association 2002.