Gaze-To-Object Mapping During Visual Search in 3D Virtual Environments

MATTHIAS BERNHARD, Vienna University of Technology
EFSTATHIOS STAVRAKIS, Cyprus Institute
MICHAEL HECHER, Vienna University of Technology
MICHAEL WIMMER, Vienna University of Technology

Stimuli obtained from highly dynamic 3D virtual environments and synchronous eye-tracking data are commonly used by algorithms that strive to correlate gaze to scene objects, a process referred to as Gaze-To-Object Mapping (GTOM). We propose to address this problem with a probabilistic approach using Bayesian inference. The desired result of the inference is a predicted probability density function (PDF) specifying for each object in the scene a probability to be attended by the user. To evaluate the quality of a predicted attention PDF, we present a methodology to assess the information value (i.e., likelihood) in the predictions of different approaches that can be used to infer object attention. To this end, we propose an experiment based on a visual search task which allows us to determine the object of attention at a certain point in time under controlled conditions. We perform this experiment with a wide range of static and dynamic visual scenes to obtain a ground-truth evaluation data set, allowing us to assess GTOM techniques in a set of 30 particularly challenging cases.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Factors*; I.3.3 [**Computer Graphics**]: Picture/Image Generation—*Viewing Algorithms*

General Terms: Algorithms,Experimentation,Human Factors

Additional Key Words and Phrases: gaze to object mapping, gaze analysis, eye-tracking, gaze processing, visual attention, object-based attention, video games, virtual environments

1. INTRODUCTION

Gaze tracking has become one of the most important tools in the study of human behavior and interaction with graphical software. Applications with 2D graphical user interfaces (GUIs) lend themselves to well-established gaze-analysis techniques, such as heat-maps or gaze-path visualizations, since the stimulus remains mostly static [Duchowski 2003]. Recent techniques, for example based on dynamic areas of interest [Papenmeier and Huff 2010; Bernhard et al. 2010], have enabled analyzing gaze behavior and visual attention in 3D graphical applications. These techniques leverage the rich information contained in the scene graph of modern video games and virtual environments in order to identify the object(s) of attention, rather than relying on recorded image sequences.

This work has been supported by the Austrian Science Fund (FWF) contract no. P23700-N23. Efstathios Stavrakis was supported by the Cypriot Research Promotion Foundation (RPF) under contract no. PROION/0311/37. Contact Author: M. Bernhard; email: matthias.bernhard@cg.tuwien.ac.at;

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

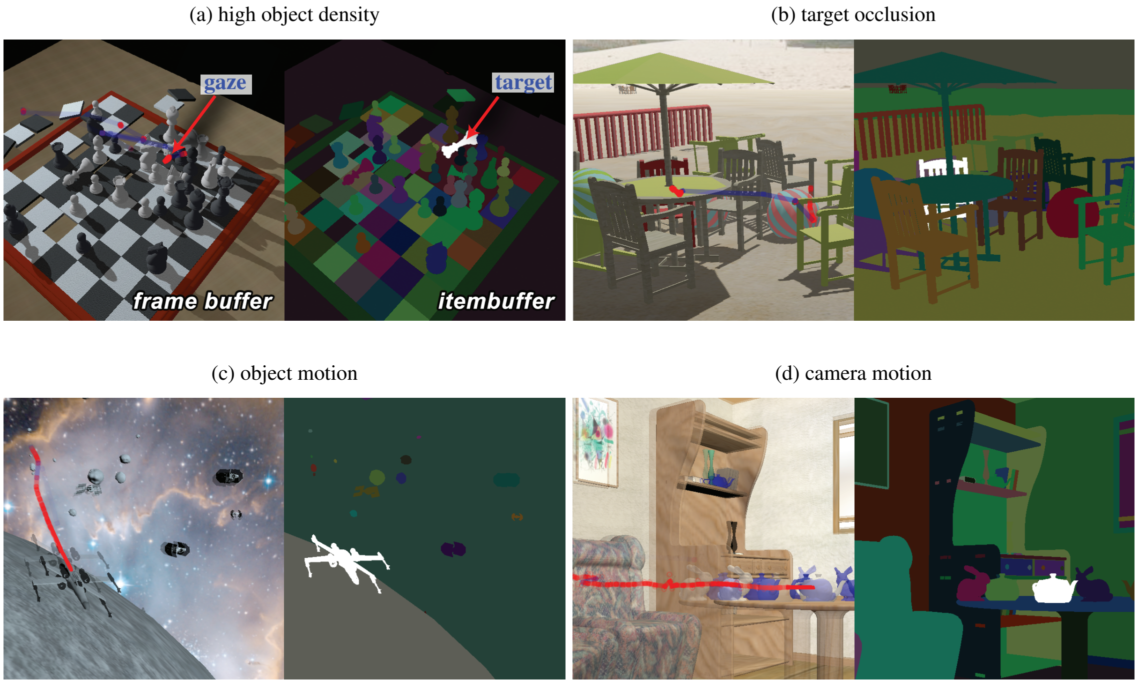


Fig. 1: Scenes exhibiting the main challenges of gaze-to-object mapping. For each scene the frame buffer (left) and the itembuffer (right) are shown. In the frame buffer the red points indicate fixations, while in the itembuffer the attention target is highlighted in white. Figure (a) is a scene with a high object density, while the target in (b) is occluded. In (c) objects are animated and in (d) the camera is in motion.

The mechanisms behind the deployment of gaze over visual stimuli is still an open research topic in visual perception, and a significant body of research suggests that attention arises from distributed interactions within and among different types of perceptual representations (e.g., spatial, featural, and object-based) [Kravitz and Behrmann 2011]. In this work, we focus on object-based attention [Duncan 1984; Huang 2010; Chen 2012] since 3D objects are those scene units that are typically sought for manipulation, interaction and representation in 3D virtual environments. Therefore, identifying objects that are (potential) attentional targets is important for a multitude of applications.

Using an eye tracker to record gaze while viewing visual stimuli generated by a 3D application (e.g., a game) is becoming a widely popular method of studying visual perception, including attention. The process of identifying the scene object that is the target of attention is called gaze-to-object mapping (GTOM). The primary challenge these algorithms face is to accurately provide information on *what* (i.e., scene object) the user is attending to, which oftentimes does not coincide with *where* the user's gaze is deployed (i.e., positions of recorded gaze). Inferring what a user is attending to from where a user is looking at is of high importance and utility, since it forms the foundation of modeling visual attention, and potentially other cognitive processes. A small number of GTOM methods geared towards dynamic 3D scenes have already been devised [Sundstedt et al. 2013], however these methods are still in their infancy and they have not been formally evaluated. Therefore, their validity and accuracy remain uncharted territory, which has motivated the more thorough evaluation presented in this work.

There are various challenging cases GTOM techniques have to deal with that are common in 3D, some of which are shown in Figure 1. For instance, due to eye-tracking inaccuracies and rapid involuntary eye movements, gaze may be deployed over multiple objects at different depths when object density is high (see Figure 1(a)). Furthermore, attentional targets can become highly occluded due to dynamic scene changes (see Figure 1(b)), and it is very frequent that scene objects or the camera are in motion, as shown in Figures 1(c) and 1(d) respectively. These cases render commonly used GTOM techniques, such as areas of interest, difficult or impractical to use in dynamic 3D scenarios.

In this work, we adopt and further assess an established methodology for gaze analysis which has been initially devised by Sundstedt et al. [Sundstedt et al. 2008] and later proposed as a gaze-analysis methodology for computer games [Sundstedt et al. 2013]. This methodology assumes that the input of any GTOM method is gaze data and an item buffer [Weghorst et al. 1984], which is an object segmentation image in which the color of each pixel encodes the unique identity of the respective object rendered to the pixel fragment. For stimuli that are computer generated, such an image can be obtained from an object-based scene graph which is commonly used by content designers to semantically structure scene geometry. Assuming a well-structured content, the item buffer provides a useful hypothesis about the perceptual organization of the scene, i.e., it specifies a set of boolean data structures which map an object identification number to a set of coherent locations which a user might perceptually group and perceive as an object. We further the understanding of the GTOM problem by proposing, formalizing and comparing several approaches. Since it turned out that GTOM is a hard problem, we decided to start with simple, intuitive and reasonable techniques that fit to our framework, which assumes that all stimulus information is represented in a sequence of item-buffer images that is passed to a tool box for gaze analysis. We have then evaluated their accuracy on a set of 30 test scenes covering a broad range of difficulty levels. We consider the desired output of a GTOM method as a predictive probability $P(\mathbf{O}|\mathbf{g})$ for each discrete object $\mathbf{O}$ to be attended by the user, given the gaze observation $\mathbf{g}$. To derive such a probability, we propose to implement GTOM methods using Bayesian inference. Results of the GTOM output can then be used to derive object space-fixation statistics, which can be further analyzed according to the semantics of the respective objects [Sundstedt et al. 2008; Sundstedt et al. 2013; Bernhard et al. 2010]. The goal is to evaluate different variants of GTOM approaches against a large set of challenging cases. For this, we introduce an experimental methodology to obtain a ground-truth data set suitable for the evaluation of GTOM methods. This is not trivial, since eye-tracking results are often considered ground truth themselves when a user's attention target has to be determined.

We believe that understanding and modeling accurate GTOM in dynamic 3D virtual environments is important because GTOM can be utilized as an implicit means of forming and verifying novel hypotheses of human visual attentional mechanisms. We provide in this work the foundation toward tackling the GTOM problem by making the following contributions:

- an introduction and assessment of several GTOM approaches for dynamic 3D stimuli, formalized via a Bayesian approach.
- an experimental methodology, which uses a set of 30 different 3D scenes, to perform an objective evaluation of GTOM methods in visually challenging cases.

2. RELATED WORK

GTOM is fundamental to eye-tracking software applications. In such applications, recorded eye gaze is first processed to identify fixations, which provide on-screen locations where gaze has been deployed by the user. Fixations are subsequently collected for elements in the scene within one frame, but may also be accumulated over time, and correlated to elements in the stimuli. According to the taxonomy proposed by Salvucci and Goldberg [Salvucci and Goldberg 2000], fixation identification algorithms can be categorized based on their spatial and temporal characteristics. Spatial fixation filters may take into account the velocity of gaze samples (velocity-based), the spread distance of fixation points (dispersion-based) or they may use an area of interest (AOI) to identify gaze samples belonging to a fixation (area-based). Temporal filters may be locally adaptive by exploiting the temporal coherence of gaze, or use a time duration threshold to account for the fact that fixations occur after 100 ms. When studying user behavior in interactive applications, especially video games, the visual signal each user perceives emerges from the pattern of his interactions and therefore differs among users. Until recently, the primary means of studying gaze behavior in VEs involved capturing all images displayed to the user while simultaneously recording his gaze [Peters and Itti 2008]. More recent work [Sundstedt et al. 2008; Stellmach et al. 2010; Bernhard et al. 2010] has turned to the use of the scene-graph data structure that the majority of 3D applications utilizes to represent all scene entities, their state and semantics. Its state can be sampled over time to obtain or infer the characteristics of any object, including their geometric extent, their screen-space projection, visible colors, motion, etc. Such a representation enables studying visual perception in object space rather than screen space.

Mapping eye gaze to stimuli requires gaze and stimuli to be captured simultaneously and subsequently processed to obtain representations suitable for GTOM algorithms. A comparison of nine GTOM methods designed primarily for gaze-controlled applications is presented in [Špakov 2011]. Two different methods implementing dynamic areas of interest are presented in [Sundstedt et al. 2013; Bernhard et al. 2010] and [Papenmeier and Huff 2010], while 3D attentional maps [Stellmach et al. 2010] and 3D attention volumes [Pfeiffer 2012] perform GTOM directly in 3D.

The most straightforward way is to cast a ray into the scene and determine the closest scene object intersected by this ray, a general-purpose computer graphics technique known as *Ray Casting*. An early example using this approach to determine attended objects in a three-dimensional scene can be found in Starker and Bolt's work [1990], where a ray shot into the scene was intersected with the scene objects' bounding spheres. Coupled with an item buffer, this strategy can be implemented by sampling the object id from the pixel fragment corresponding to the current gaze position. This is computationally inexpensive, particularly in 3D applications where geometry information resides on the GPU, but its major drawback is that it determines an object as the attention target with a binary test (i.e., is intersected or not). The problem stems from the fact that gaze locations sensed by eye trackers have limited accuracy and, more importantly, users do not necessarily center gaze directly on pixels belonging to the attended object. Therefore, ray casting is not necessarily a reliable method to infer the attended object. To mitigate the problem of determining objects at a single position, Sundstedt et al. [2008] proposed to sample the neighborhood of the gaze position in an item buffer using a Gaussian splat corresponding to the size of the fovea as a weighting function. This strategy yields an importance value which can be used as a probability prediction. An alternative to rendering an item buffer, proposed recently by Mantiuk et al. [2013], is to sparsely attach target points to object surfaces and use those to compute the distance between gaze and these target points. They further proposed to utilize motion information to increase accuracy, since there is a certain amount of correlation between gaze movement trajectories and the motion of an attended target point, which can be transformed to a probability estimate. Mapping gaze to target points placed on scene geometry enabled them to determine the depth of the features currently attended in order to control the depth-of-field according to the attention of the user. Moreover, they processed this probability over time with a Hidden Markov Chain, which served as a fixation filter and guaranteed temporal coherence in the selection of the current depth-of-field distance.

3. PIPELINE

Mapping gaze to objects in 3D scenes can be considered as the process of computing the spatio-temporal correlation of gaze information to objects in a 3D scene. The process can be broken down into the following three steps:

Gaze and Stimulus Processing: Raw gaze data are processed to obtain a smooth gaze signal, which is filtered to detect fixations.

Spatial GTOM: Fixations are used to compute an object-space probability density distribution for each frame in the stimuli.

Temporal GTOM: A probability over many frames (e.g., a time window or an entire fixation) is computed using an accumulation strategy of spatial GTOM probabilities.

3.1 Gaze and Stimulus Processing

GTOM methods take the following temporal signals as input: a smooth gaze position $g(t)$, a fixation signal $f(t)$, which is a sequence of bits denoting whether a user fixates at time t, and a stimulus signal $s(t)$, which contains information about the objects in the scene. To obtain these signals, we perform the following three processing steps:

Gaze Filtering (provides $g(t)$). A spatio-temporal low-pass filter is used to process noisy raw gaze data into a smooth gaze data signal. This signal encodes a 2D gaze position that estimates the center between the gaze of the left and the right eye of the user. To this end, we used a bilateral low-pass filter [Tomasi and Manduchi 1998] on the gaze signal. We use a Gaussian filter kernel with a standard deviation of $\sigma_t = 0.5sec$ for the temporal kernel and $\sigma_s = 0.7°$ for the spatial kernel, respectively. The advantage of bilateral filtering is that it reduces over-smoothing over the abrupt transitions between fixations and saccadic gaze movements. The noise-suppressed signal is used for two purposes: first, to increase the robustness of subsequent saccade and fixation identification, and second to estimate the fixation position, which we assume to be a function of time.

Identifying Fixation Times (provides $f(t)$). Based on the assumption that attention correlates with fixation locations only [Duchowski 2003], we identify the fixation state (i.e., fixating or not) to partition gaze data into different fixations. We found that dispersion filters often used for fixation identification are less appropriate, as dispersion increases proportionally to drift velocity, and therefore, single moving fixations, such as those occurring when tracking a moving object, are identified as a series of shorter fixations. Instead we utilized a velocity-based fixation detection method that avoids this problem. We found an acceleration threshold-based saccade detector, as proposed by Tole and Young [1981], to be most appropriate since it performs best in identifying smooth pursuits, which frequently occur in dynamic stimuli. This detector identifies saccades by their ballistic properties, which have a strong acceleration ($> 40000°/sec^2$) at the onset and a deceleration of similar magnitude at the end of a saccade. Fixations are then determined as groups of non-saccadic consecutive gaze points with a minimum duration (e.g., 3 gaze samples).

Stimulus Processing (provides $s(t)$). Each 3D object is allocated a unique 24-bit object id, and its projection on the 2D viewing plane is colored with the RGB value corresponding to this object id, which results in an item buffer [Weghorst et al. 1984]. Computing an item buffer is described in more detail in [Sundstedt et al. 2008] and [Bernhard et al. 2010].

3.2 Gaze-to-Object Mapping (GTOM)

After preprocessing gaze and stimulus data, we map gaze to objects for each frame during a fixation. The desired output of a GTOM strategy is an object-space probability density function (PDF) for a discrete moment in time (t). This can be expressed as the posterior probability $P^{(t)}(\mathbf{O}|\mathbf{g})$ for object $\mathbf{O}$ being attended by the user at time t, given gaze information $\mathbf{g}$. The required information, which we use to infer the object of attention, is extracted from the stimulus signal $s(t)$.

Note that this work focuses on a thorough evaluation of *per frame* GTOM results. That is, our evaluation was aimed at the problem of spatial gaze-to-object mapping in each frame without utilizing information from results of previous frames. Hence, we followed a simple strategy to process GTOM results over time by accumulating the probabilities obtained in each frame over the time window of an entire fixation.

4. GTOM APPROACHES

In this section, we describe several methods for mapping gaze to objects in individual frames, which are all encapsulated in a formal Bayesian inference framework. We define two categories of gaze-to-object mapping techniques: *Passive Attention* GTOM (Section 4.1) and *Active Attention* GTOM (Section 4.2), illustrated in Figure 2 as simple graphical models. In both cases, the observed variable is the gaze position $\mathbf{g}$, while a probability is inferred for the hidden variable $\mathbf{O}$.

In *Passive* GTOM, we consider mapping gaze to features (e.g., pixels or target points). The hypothesis is that object attention follows gaze, i.e., where a user looks determines which features are seen and hence attended. A problem of this strategy is that attention is evaluated only locally near gaze positions without taking into account that other (more distant) parts of the objects, or even other cognitive processes such as memory [Hollingworth 2012], may have influenced gaze programming. Therefore, in *Passive* GTOM a bias toward selecting objects as the attentional targets that occupy larger portions of the stimuli is likely. If gaze positions are distributed randomly in space, variations in size cause a stochastic advantage for larger objects to be inferred as attention targets.

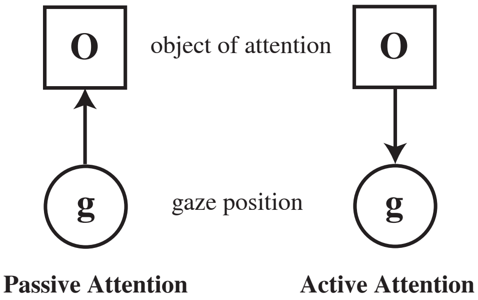


Fig. 2: Simple graphical model to illustrate the two different approaches one can use for inferring attended objects from gaze. A square is used to illustrate that the object of attention is a discrete variable (e.g., an ID), while a circle is used to indicate that gaze is a continuous variable.

An alternative hypothesis is that object attention determines gaze ($\mathbf{O} \rightarrow \mathbf{g}$), which we refer to as *Active* GTOM. *Active* GTOM techniques specify an estimate for the posterior probability $P(\mathbf{g}|\mathbf{O})$ that gaze $\mathbf{g}$ follows given that object $\mathbf{O}$ is attended. This approach assumes that attention is equally distributed in object space, i.e., there is no stochastic advantage for large objects. Only high-level factors, such as task relevance, can be used as priors to increase the probability of particular objects being attended before gaze has been observed. Moreover, this approach accounts for situations where a user pays attention to a particular object as a whole and not to specific features. In this case, all features of the object could contribute holistically to the selection of an optimal viewing position.

4.1 Passive GTOM – *Gaze Determines Attention*

In *Passive* GTOM, we assume that the gaze $\mathbf{g}$ determines what is seen with foveal vision and thus attended. To estimate a probability for an object to be attended, we assume that attention activates object features near the current gaze position more than distant features. We use the amount of activation estimated for features $\mathbf{o} \in \mathbf{O}$ being located on object $\mathbf{O}$ to derive a probability $P(\mathbf{O}|\mathbf{g})$ that object $\mathbf{O}$ is attended given that gaze $\mathbf{g}$ was observed. To avoid normalization issues, we estimate a proportional function $P(\mathbf{O}, \mathbf{g})$, from which we derive a posterior probability by normalization:

$$P(\mathbf{O}|\mathbf{g}) \propto P(\mathbf{O}, \mathbf{g}) \tag{1}$$

Ray Casting (RC). The simplest approach to estimate $P(\mathbf{g}, \mathbf{O})$ is to perform a binary evaluation whether a gaze position $\mathbf{g}$ is located on object $\mathbf{O}$ or not:

$$P^{RC}(\mathbf{O}, \mathbf{g}) = \begin{cases} 1 & \text{if } \mathbf{g} \in \mathbf{O} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Closest Feature Mapping (CFM). Another option is to assume that a user attends to the closest feature $\mathbf{o}$ of an object $\mathbf{O}$ and weight the euclidean distance with a Gaussian with a standard deviation of θ_s:

$$P^{CFM}(\mathbf{O}, \mathbf{g}) \propto \exp\left(\frac{\min\limits_{\mathbf{o} \in \mathbf{O}} \left\{ \|\mathbf{o} - \mathbf{g}\|^2 \right\}}{-2\theta_s^2}\right) \tag{3}$$

In theory, this strategy requires looking-up every pixel in the item buffer, i.e., complexity is $\mathcal{O}(N)$, with N denoting the number of pixels in the item buffer. However, since the energy of pixels exponentially drops to zero as the distance increases from the gaze location, a clipping radius (e.g., $r = 4\theta_s$) can be used to obtain an approximate solution without a significant loss of accuracy. Furthermore, in realtime applications a portion of the item buffer, which is the bounding window around the clipping circle, may be used to increase processing speed.

Fovea Splatting (FS). A second alternative is to assume that a user attends to multiple features at the same time and to infer the attention probability by accumulating all features/pixels $\mathbf{o} \in \mathbf{O}$ weighted by a Gaussian splat of the size of the fovea [Sundstedt et al. 2008] (with standard deviation θ_s):

$$P^{FS}(\mathbf{O}, \mathbf{g}) \propto \sum_{\mathbf{o} \in \mathbf{O}} \exp\left(\frac{\|\mathbf{o} - \mathbf{g}\|^2}{-2\theta_s^2}\right) \tag{4}$$

The computational complexity of this method is equivalent to Closest Feature Mapping. However, instead of comparing distances while searching the closest pixel of each object we compute the Gaussian energy at each pixel and accumulate these values. Computing the energies can be optimized by using a pre-computed kernel.

4.2 Active GTOM – *Attention Determines Gaze*

In *Active* GTOM approaches, we assume that once an attention target is selected, gaze is directed towards it. This model estimates $P(\mathbf{g}|\mathbf{O})$, which is the posterior probability that $\mathbf{g}$ is observed given that $\mathbf{O}$ is attended. To infer the probability that $\mathbf{O}$ is attended from $\mathbf{g}$, we apply the Bayes Theorem, and thus we also call *Active* GTOM methods "Bayesian Inference GTOM" (BIGTOM) approaches:

$$P(\mathbf{O}|\mathbf{g}) = \frac{P(\mathbf{g}|\mathbf{O})P(\mathbf{O})}{\sum_i P(\mathbf{g}|\mathbf{O}_i)P(\mathbf{O}_i)} \tag{5}$$

The prior $P(\mathbf{O})$ can be used to increase the probability of task-relevant objects or to propagate previous GTOM results over time. However, our research focuses on one frame without any *a priori* knowledge, and thus we use a constant, which is theoretically $P(\mathbf{O}) = \frac{1}{N}$ (where N is the number of objects), in our evaluation.

Concerning the influence of attention on gaze position, we will investigate different ways of estimating the posterior $P(\mathbf{g}|\mathbf{O})$:

Center of Gravity Mapping (CM). A straightforward, but effective strategy, is to exploit the strong tendency of users to focus at the center of an object [Henderson 1993]. Similar to [Xu et al. 2008] (for mapping to words in a text documents) and [Mantiuk et al. 2013](for mapping to small objects), we place a monovariate, but two-dimensional, Gaussian at the center of gravity of the object with a kernel size θ_s:

$$P^{CM}(\mathbf{g}|\mathbf{O}) = \frac{1}{2\pi\theta_s^2} \exp\left(\frac{\|\mathbf{g} - \mathbf{c_O}\|^2}{2\theta_s^2}\right) \tag{6}$$

To compute the center of gravity, we extract the set of (x, y)-coordinates of the pixels belonging to the respective object $\mathbf{O}$ and compute the average position of these samples.

In theory, this strategy requires computing the center of gravity for each object in the scene, i.e., complexity is in $\mathcal{O}(N)$. However, objects which are distant from the gaze position can be disregarded by assuming that their attention probability is zero. To this end, we can determine which objects have at least one pixel in the vicinity of the gaze position in a circular region around $\mathbf{g}$ (e.g., $r = 4\sigma_s$). The center of gravity is then determined for those objects only, while the rest have a zero probability. Another acceptable optimization is to use sparse sampling (e.g., $stepsize = 3pixels$), since the center of gravity is relatively stable against this simplification. Nevertheless, in contrast to *passive* GTOM approaches, it is recommended to generate and copy a full-screen item buffer, where resolution can be eventually down-sampled according to the step size used for sparse sampling. A coarse simplification, which could be considered in applications where an item buffer is difficult to obtain, is to use the center of the object's bounding box projected to device-space coordinates. However, accurate visibility information is not available and thus should also be estimated, e.g., by using the depth of the world space bounding box. This is necessary to exclude objects which are barely visible, or to trim bounding boxes such that they exclude occluded object parts.

Normalized Closest Feature Mapping/Fovea Splatting (nCFM / nFS). Since the strategy described above assumes that a user targets gaze near the center of an object, we will also investigate strategies which better account for the object's projected shape (i.e., the pixel coverage in the item buffer). We follow the CFM proposed for *Passive*-GTOM approaches, but normalize as follows:

$$P^{nCFM}(\mathbf{g}|\mathbf{O}) = \frac{P^{CFM}(\mathbf{g},\mathbf{O})}{\int P^{CFM}(\mathbf{x},\mathbf{O})d\mathbf{x}} \tag{7}$$

Here, we modified the closest feature-mapping method by specifying a function which sums up to 1.0 for all possible gaze positions $\mathbf{x}$. Analogously, we normalize the fovea splatting method:

$$P^{nFS}(\mathbf{g}|\mathbf{O}) = \frac{P^{FS}(\mathbf{g},\mathbf{O})}{\int P^{FS}(\mathbf{x},\mathbf{O})d\mathbf{x}} \tag{8}$$

Both methods have in theory a computational complexity of $\mathcal{O}(N^2)$ since they correspond to a convolution. Simplifications that could be performed are: (a) evaluating only objects which are close to the gaze position, as described for CM, (b) sparse sampling of the item buffer, and (c) truncation of the splat kernel with a clipping circle.

4.3 Utilizing Motion Information

During smooth pursuits, we assume that the visual system attempts to minimize the velocity of an object relative to the gaze movement in order to maintain a stable retinal image. We utilize this behavior to improve the accuracy of the inference. To this end, we compute a retinal stability probability P^{RS} by using the gaze movement vector $\dot{\mathbf{g}}$ and object movement vector $\mathbf{m_O}$:

$$P^{RS}(\dot{\mathbf{g}},\mathbf{O}) \propto \exp\left(\frac{\|\dot{\mathbf{g}} - \mathbf{m_O}\|^2}{-2\theta_r^2}\right) \tag{9}$$

To model the standard deviation between feature and gaze motion, a retinal drift tolerance parameter θ_r is introduced. To obtain gaze movement information, we differentiate the filtered gaze position along consecutive gaze samples. To further smooth the gaze motion samples in fixations, we use a Gaussian filter kernel with a standard deviation corresponding to 2 gaze samples (sampled with $50Hz$). Object motion vectors are estimated by differentiating the temporal signal of gaze and object center positions ($\mathbf{c_O}(t)$).

We combine this probability with the position PDF's specified above by a multiplication:

$$P(\mathbf{O}|\mathbf{g},\dot{\mathbf{g}}) \propto P(\mathbf{O}|\mathbf{g})P^{RS}(\mathbf{O},\dot{\mathbf{g}}) \tag{10}$$

4.4 Relation to Previous Work

Three of the methods presented in this Section are related or equivalent to solutions proposed in previous work. These are Ray Casting (e.g., [Starker and Bolt 1990]), Fovea Splatting ([Sundstedt et al. 2008]) and Center of gravity Mapping (e.g., [Xu et al. 2008]). Note also that Ray Casting and Center Of Gravity Mapping could not immediately be used in our framework and had to be adjusted for the use with an item buffer. Mantiuk et al. [Mantiuk et al. 2013] has introduced the idea of using motion information together with Gaussian distance metrics.

The other three methods (i.e., Closest Feature Mapping, normalized Closest Feature Mapping/Fovea Splatting) are, to the best of our knowledge, first proposed in this work.

5. EVALUATION METHODOLOGY

In this section, we describe the eye-tracking experiment and the scenes used to obtain ground-truth data samples necessary for the evaluation of GTOM methods. Moreover, we will explain how the obtained data set is used to evaluate the accuracy of GTOM methods (Section 5.4).

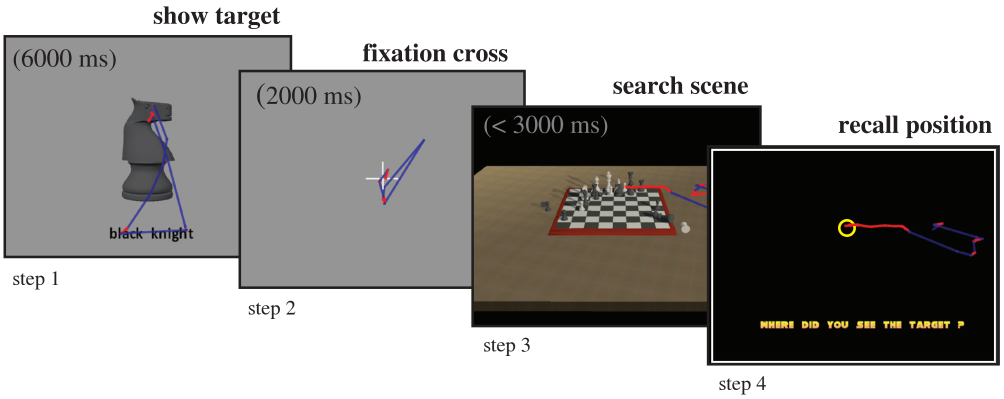


Fig. 3: Procedure of one trial (in this case target-present). In step 1, the participant previews a target; in step 2 fixates on a cross (for validating later the eye-tracking accuracy); in step 3, he searches for the target. In step 4, the participant recalls the position of the target on a blank screen to ensure that he did not guess.

5.1 Background

To evaluate and compare the accuracy of GTOM methods, we need ground-truth data comprising pairs of known fixations ($\mathbf{F}_i$) and known attentional targets ($\mathbf{T}_i$). However, it is not trivial to find the actual attentional target $\mathbf{T}_i$, since the common way to do so is to use eye tracking and apply GTOM, the method we want to improve in the first place.

To solve this problem, we designed an eye-tracking experiment where we direct the user to attend a particular object (i.e., we prescribe $\mathbf{T}_i$) and use eye tracking to gather the corresponding fixation $\mathbf{F}_i$. The task we use for that purpose is a visual search task, because this does not hinder the user's ability to naturally deploy his attention in a scene, and requires active deployment of attention to locate a target. Thus, a visual search task can be used to *implicitly* instruct participants to fixate on a particular object in a scene, without visually altering or annotating the objects. According to [Hollingworth 2012], visual memory facilitates guiding attention and gaze during visual search, therefore we provide participants with a visual preview and an on-screen description of each target before each search task. To obtain an objective evaluation as possible, our experimental test bed includes an extensive set of 30 different scenarios that arise frequently in visual scenes (e.g., video games).

5.2 Eye-Tracking Experiment

Setup and Participants. We used a Tobii x50 eye tracker (50 Hz), which was placed in front of the display of an Intel Core 2 Duo workstation with a 2.4 GHz CPU, 2GB RAM and an NVIDIA GeForce 8800 GTX graphics card. This setup was sufficiently powerful to simultaneously run the application at a frame rate greater than 50 fps, and also operate the eye tracker. The display was a commodity LCD display (IBM ThinkVision L200p with a resolution of 1600×1200 pixels at 100 dpi). 28 subjects, with ages between 20 and 55, participated in the study (12 females). All subjects had normal or corrected-to-normal vision. Participants were seated comfortably in front of the screen's center at a distance of 60 cm. No chin rest was used so as to allow for natural viewing behavior. By manually inspecting the recorded gaze data using our visualization tool, we found that only 22 gaze data sets were of reliable accuracy to be included in the evaluation data set. The other 6 data sets were all from participants wearing contact lenses and exhibited serious systematic errors in the gaze recordings, which became most apparent during smooth-pursuit object tracking and when the fixation cross was shown.

Task and Procedure. Each trial consisted of 4 steps: (i) preview a target object (6000 ms), (ii) fixate on a cross (2000 ms), (iii) search a scene for the target (< 3000 ms), and (iv) recall the position of the target. This procedure is shown in Figure 3 and resembles experiments commonly used to verify attention models by measuring reaction times

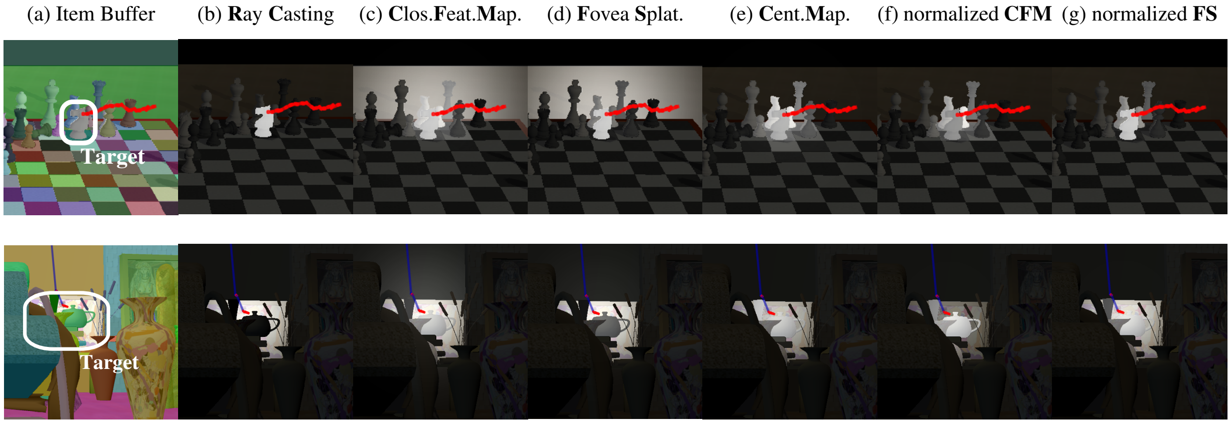


Fig. 4: Visualization of the item buffer and the predicted PDFs for each method. In the chess scene the camera was moving in the right direction and the attention target was the knight figure. In the scene shown in the bottom row, the search target was the painting located in the background of many occluding foreground objects.

in visual search tasks (e.g., [Wolfe 2000]). Experiment participants performed a block of 50 trials. The participants responded by pressing a key in case they located the target while in step (iii), otherwise they were instructed not to act. To motivate quick responses in all search trials, participants were told that their reaction time is measured. Among the 50 trials, 30 were target-present and 20 target-absent trials. In target-absent trials, participants had to withhold the response and wait until the application automatically started the next trial after reaching a timeout of 2000 ms in static scenes or after seeing one animation cycle in dynamic scenes. Prior to the main trials, each participant practiced the procedure on a block of another 10 trials (with different scenes). The eye tracker was calibrated between the practice and main block. Using the software provided by the manufacturer, a calibration was calculated based on 5 control points.

Obtaining the Ground Truth. In order to make sure fixations really correspond to specific attended objects, we need to exclude detection methods other than eye fixation, eliminate guessing, and perform verification.

In each scene, we placed distractor objects sharing a feature (e.g., color, orientation or shape) with the target in at least one dimension. This causes targets and distractors to be indistinguishable by pre-attentive selection [Treisman and Gelade 1980], and thus compels a user to identify targets by performing a serial search using eye fixations. From each target-present search trial, we obtain the ground-truth attention target and a respective fixation.

Through offline manual inspection, we selected these fixations using the following rules: **(a)** the fixation started briefly before a participant's response, **(b)** the fixation is reasonably close ($< 2°$) to the target, and **(c)** a participant could correctly recall the target position. To check (c), participants had to place after each target-present search trial a circle on the target's last position within a void black screen (Figure 3).

From these strategically selected fixations, we generated an evaluation data set in which each entry consists of the target object's identification color code and the time-stamped gaze data, as well as the series of time-stamped item buffer images of the time window corresponding to this fixation.

5.3 Scenes

To highlight particular strengths and weaknesses of the evaluated algorithms, we designed a total of 30 different scenes featuring a variety of levels of difficulty for GTOM. We categorized 27 of these scenes into 2 groups: **(1)** 9 scenes without camera or object animation (group **Static**), and **(2)** 18 scenes where the target or the camera, or both, are

Method	θ_s	θ_r	max LLH
Ray **C**asting	-	-	-833
Closest **F**eature **M**apping	$0.7°$	$12°/s$	-691
Fovea **S**platting	$0.7°$	$12°/s$	-726
Center of Grav.**M**apping	$1.0°$	$12°/s$	-481
normalized **CFM**	$0.7°$	$12°/s$	-565
normalized **FS**	$0.7°$	$12°/s$	-419

Table I. : Optimal parameters and the maximum log-likelihood of each method.

animated (group **Dynamic**) . In each group, we used different target sizes and object densities, as well as varying degrees of target occlusion. To reveal a potential bias of the method towards foreground scene objects, we constructed cases where a user has to search for a background object. Thus, the 3rd scene group, which we denote **Background**, includes 3 scenes where the target is an abstract painting and located in the background of many occluding foreground objects. One scene was static, the second was shown from a moving viewpoint and in the third, occluding foreground objects were animated.

All scenes were created and animated in Maya and exported with OgreMax to an XML file (including animation keyframes) and a set of mesh files to be loaded with the Ogre3D rendering engine that we used to build the experiment's application.

5.4 Performance Measurement and Parameterization

The data sets collected through the experiment presented in Section 5.2 will be used to evaluate and measure the accuracy of the GTOM methods proposed in Section 4. As a quantitative performance metric, we compute the log-likelihood (LLH) by summing the logarithm of the predicted probability for each evaluation data set:

$$LLH(\theta, M) = \sum_i ln \max \{P(\mathbf{O} = \mathbf{T}_i | \mathbf{F}_i, \theta, M), \epsilon\}, \tag{11}$$

where $P(\mathbf{O} = \mathbf{T}_i | \mathbf{F}_i, \theta, M)$ specifies the probability predicted by a GTOM method M, configured with the parameter set θ, for the ground truth attention target $\mathbf{T}_i$. A small threshold $\epsilon = 0.01$ was used to remove outliers (i.e., zero-probability predictions, which may cause an infinite LLH, e.g. when $M =$ RC). We searched for the maximum LLH to specify the optimal parameters for each method which are listed in Table I.

In addition, we use the success rate, which is a simple ordinal measure. We compute it for each fixation by the proportion of frames where a GTOM method has predicted the highest probability for the ground truth target.

6. RESULTS AND DISCUSSION

An example visualization of the GTOM results for different methods is shown in Figure 4 and in the accompanying video. The visualizations demonstrate that depending on the method and visual scenario, the predicted PDFs can vary considerably.

Two measures are used to estimate each algorithm's performance in predicting the object attention probabilities, as described in Section 5.4: (a) the average likelihood (normalized by the number of fixations), and (b) the success rate. In Figure 5 we depicted the results for the success rate and likelihoods. To illustrate the variation among subjects, we computed both measures by grouping fixations of all scenes of a category (Static, Dynamic or Background) and then ranked the results for each participant (Figure 5a). To show the variation of performance among different scenes, we grouped fixations of each participant for all scenes of a category and ranked the scores for each scene (Figure 5b). To provide a more intuitive illustration of the likelihood scores, we used in these graphs the exponential of the log-likelihood scores which were normalized by the number of fixations being grouped. We call this measure *average likelihood* and it corresponds to the expected likelihood that a certain target object and fixation pair ($\{\mathbf{T}_i, \mathbf{F}_i\}$) is observed, assuming that a particular GTOM model is correct.

In Figure 5b it can be also seen that the test scenes cover a broad range of difficulty levels, since the success rate and average likelihood drop smoothly from high to low scores for most methods. Also there is a considerable variation across different participants. An ANOVA on both measures (Model: $score \sim method + scene * partipant$) proved that there is a highly significant effect of the factors $method$ ($F_{5,2696} \geq 78.34, p < 10^{-15}, \eta^2 \geq 0.07$/ moderate effect size), $scene$ ($F_{29,2696} \geq 41.08, p < 10^{-15}, \eta^2 \geq 0.24$ / large effect) and $participant$ ($F_{21,2696} \geq 3.48, p < 10^{-6}, \eta^2 \geq 0.013$/ small effect), as well as in the interaction between $participant$ and $scene$ ($F_{488,2696} \geq 2.40, p < 10^{-15}, \eta^2 \geq 0.20$ / large effect). Thus, we evaluated the significance of differences in a post-hoc analysis of the LLH-scores and success rates by using a generalized mixed-effect model (R-function *lmer* from package *lme4*), which is suited to fit to our repeated measurements data with the two interacting random effects $scene$ and $participant$ (Model: $score \sim method + (1|scene/partipant)$). On this model, we applied a Tukey-HSD test (using the R-function *glht* from package *multcomp*). A visualization of the HSD-test results can be found in Figure 6.

Comparing the different GTOM methods, the overall results suggest that *active* GTOM methods outperform *passive* GTOM approaches in the scenes where the target is a compact object.

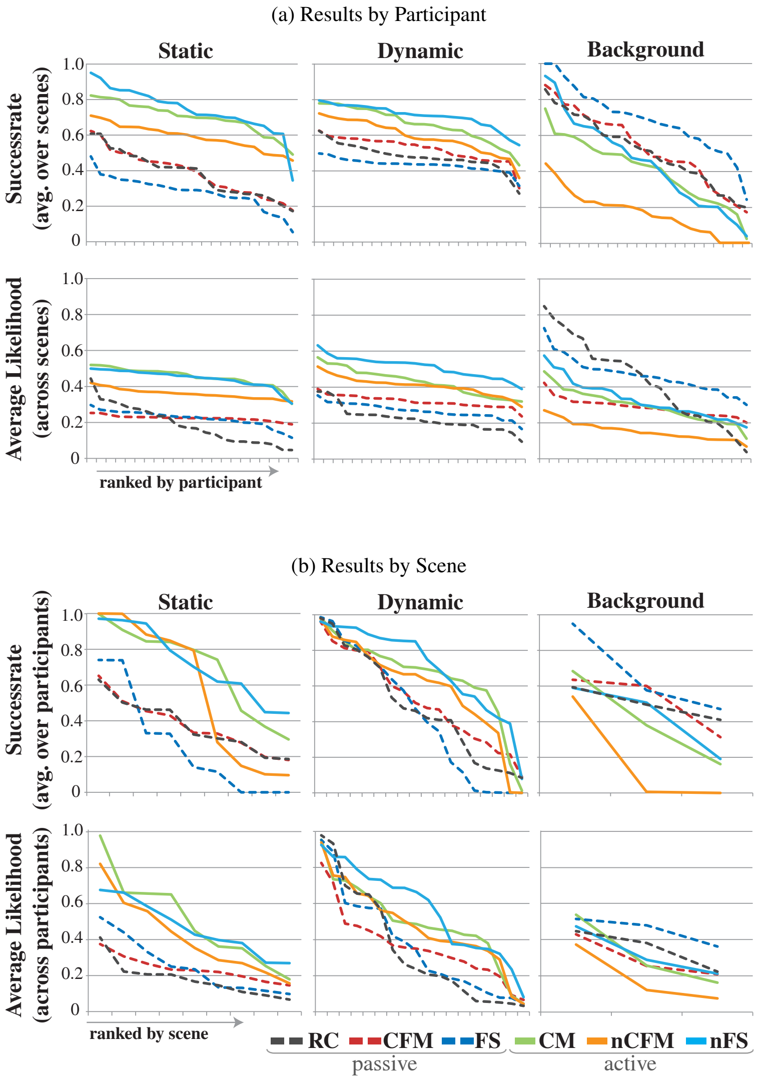
Passive GTOM approaches are more conservative as they predict attention to those features which are actually close to the center of the fovea. Moreover, they infer attention locally from object features near the current gaze position and thus are also less computationally expensive. *Passive* GTOM methods provide reliable probability predictions in cases where a user focuses on object details or on objects where features dominate (i.e., background objects). Other cases where *passive* methods provide good results are scenes with objects of similar size and few occlusions. However, since the spatial stochastic advantage of large objects is not compensated for *passive* approaches, those tend to increase importance of objects covering many pixels on the screen. This overestimation may become particularly severe with the FS approach. When compared to CFM or RC, FS performs poorly since the predicted probability increases for surrounding large objects due to integrating the energy of many pixels around the current gaze position.

Thus, when objects vary considerably in size or occlude each other, *active* methods provide better results. In contrast to *passive* methods *active* approaches treat *a priori* each object equally important, independent of the number of pixels covered on the screen. Besides compensating for the spatial stochastic advantage, *active* GTOM approaches better account for situations where a user does not focus on object details, but rather perceives the attended object as a coherent entity. The optimal position to obtain a coherent percept of an object is the center of gravity, which turned out to be one of the most effective and robust predictors of a gaze position, assuming the object of attention is known. Similar observations have also been obtained in experimental psychology research (e.g., [Henderson 1993]).

Overall, center-of-gravity-based methods appear to provide good results in our evaluation, particularly in cases where a target is occluded. The simplicity of CM is a great advantage for many applications and can be easily implemented by just using screen-space bounding boxes. Surprisingly, CM also performs well in scenes where a background object had to be searched. However, one reason for this could lie in our choice of objects to be compact, to some extent, in order to enable a meaningful visual search task. This is a limitation of our evaluation data set and for future work it is important to further extend our evaluation data sets with more scenarios where a user attends to background objects, such as a region in the sky or a wall. Since a user does not necessarily focus on the center of gravity of an object in this case, using the center of gravity as the only feature being evaluated introduces a strong assumption about gaze behavior which is in conflict with situations where a user focuses on object details or significant features (e.g., the handle of a teapot). nFS is a more general approach that is amongst the methods with the best scores in our evaluation and provides a significantly better probability estimate than CM in dynamic scenes. It works more consistently for a larger variety of object shapes. By integrating features around the gaze center, the circular nature of the monovariate 2D-Gaussian also yields a maximum in $P(\mathbf{g}|\mathbf{O})$ at locations near the object's center of gravity. Since this property of enhancing the probability in the center of gravity is not shared by nCFM, that method performs particularly worse.

7. CONCLUSION AND FUTURE WORK

In this work, we systematically investigated several alternatives to infer object-space attention from gaze in dynamic 3D environments. These methods are particularly important because the screen-space location where the gaze of a user is deployed does not always correspond spatially to attended objects in the stimuli. The Bayesian formulation proposed in this work aims at providing a framework within which gaze-to-object mapping techniques can be described, evalu-


(a) Results by Participant

(b) Results by Scene

Fig. 5: Success rate and average Likelihood ranked by participant (top), or scene (bottom), respectively. We compare the methods Ray Casting (RC), Closest Feature Mapping (CFM), Fovea Splatting (FS), Center-of-gravity Mapping (CM), and normalized Closest Feature Mapping/Fovea Splatting (nCFM/nFS). The results were grouped into static scenes (no motion), dynamic scenes (object or camera motion) and background scenes (large object occluded by foreground objects).

ated and compared. The resulting probabilities provide a confidence measure for gaze-based object selections or could be used for a probabilistic optimization of rendering methods (e.g., minimization of the expected attention error).

We performed a formal evaluation of methods based on a visual search task in a wide range of different 3D scenes, from which stimuli, gaze data and the corresponding ground-truth attention target were obtained. It turned out that it is favorable to use Active GTOM approaches that assume gaze follows attention. The main advantage of these methods is that they make the assumption that the basic units of attention are objects, and thus their likelihood to be attended

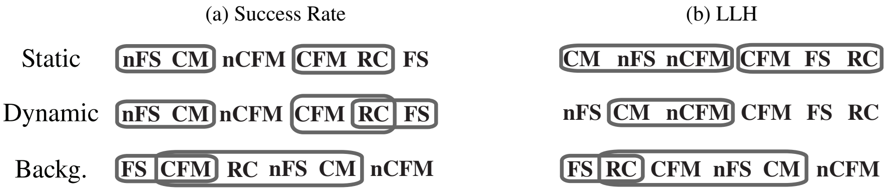


Fig. 6: Post-hoc analysis for success rate and LLH scores. Algorithms are sorted by their intercept. Those which are encompassed by a rectangle are not significantly different, i.e., they are within in a 95% statistical similarity group (Tukey contrasts).

is independent of the amount of features they bear or the number of pixels they cover on the screen at any time. Of course, in many cases, *passive* GTOM approaches also yield accurate results, when *Active* GTOM methods performed worse. Our intuition is that the two approaches are not mutually exclusive – instead, we believe they account for different attentional mechanisms (i.e., feature-oriented and object-oriented attention), which are both present under natural conditions. Thus, one direction to further GTOM methods could be to combine or unify *Passive* and *Active* approaches.

Another important challenge is to evaluate different approaches to combine probabilities over time. In this work we have only averaged over fixations, but accuracy could potentially be increased by propagating probabilities over subsequent frames, or fixations. This could be done, for instance, with Hidden Markov Chains (cf. [Mantiuk et al. 2013]). However, the choice of the temporal processing method is also a matter of the application where GTOM is used. For instance, for gaze pointing tasks, non-probabilistic approaches to accumulate GTOM results over time, such as force physics [Zhang et al. 2008], could be more adequate.

Another important step which future work should address is the way segmentation images (i.e., the item buffer) are obtained. In this work, we rely on the assumption that the way scene geometries are grouped together (by the content artist who created the virtual environment) corresponds well to the combination of features a user's visual perception groups into objects. However, even if the scene geometry is structured well into objects, there are problematic geometries, such as the terrain of a scene, or vegetation, which are difficult to cluster into perceptual objects. Thus, it would be useful to investigate alternative item buffer generation methods which further subdivide geometry by geometric or texture feature analysis. We also believe that an optimal GTOM method should be based on a hierarchical representation, which would allow inferring object attention at different levels of this hierarchy. A user may pay attention to a group of objects, a single object, to parts of an object or to particular features. The ultimate goal is, however, to utilize GTOM approaches to non-artificial stimuli or stimuli where an object segmentation is simply not available. For these applications, gaze processing has to be combined with vision algorithms (e.g., proto-object maps [Yanulevskaya et al. 2013]) that can automatically segment parts in an image which a user perceives as objects.

We assessed in this work the quantitative information value in the predicted probabilities. These attention probabilities distributed over multiple objects may find utility in algorithms that require a distribution of importances in order to optimally adjust their parameters to achieve their goals (e.g., minimization of the expected perceived error).

Improved GTOM can have a crucial impact within HCI and Computer Graphics, including attention visualization, gaze analysis, gaze-based semantic pointing, level-of-detail rendering, depth-of-field simulations, attention-aware stereo rendering, and many others.

Acknowledgements

The authors would like to thank Kartik Asooja for his assistance in creating the virtual environment scenes.

REFERENCES

BERNHARD, M., STAVRAKIS, E., AND WIMMER, M. 2010. An empirical pipeline to derive gaze prediction heuristics for 3D action games. *ACM Transactions on Applied Perception (TAP) 8,* 1, 4:1–4:30.

CHEN, Z. 2012. Object-based attention: a tutorial review. *Attention, perception & psychophysics 74,* 5, 784–802.

DUCHOWSKI, A. T. 2003. *Eye tracking methodology: Theory and practice*. Springer, New York.

DUNCAN, J. 1984. Selective attention and the organization of visual information. *Journal of experimental psychology. General 113,* 4, 501–517.

HENDERSON, J. 1993. Eye movement control during visual object processing: effects of initial fixation position and semantic constraint. *Canadian Journal of Experimental Psychology 27,* 1, 79–98.

HOLLINGWORTH, A. 2012. Task specificity and the influence of memory on visual search: Comment on Võ and Wolfe (2012). *Journal of Experimental Psychology: Human Perception and Performance 38,* 6, 1596–1603.

HUANG, L. 2010. What is the unit of visual attention? object for selection, but boolean map for access. *J Exp Psychol Gen 139,* 1, 162–79.

KRAVITZ, D. J. AND BEHRMANN, M. 2011. Space-, object-, and feature-based attention interact to organize visual scenes. *Attention, perception & psychophysics 73,* 8, 2434–2447.

MANTIUK, R., BAZYLUK, B., AND MANTIUK, R. K. 2013. Gaze-driven object tracking for real time rendering. *Computer Graphics Forum 32,* 2pt2, 163–173.

PAPENMEIER, F. AND HUFF, M. 2010. DynAOI: a tool for matching eye-movement data with dynamic areas of interest in animations and movies. *Behavior Research Methods 42,* 1, 179–187.

PETERS, R. J. AND ITTI, L. 2008. Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception 5,* 2, 1–19.

PFEIFFER, T. 2012. Measuring and visualizing attention in space with 3D attention volumes. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM Press, 29–36.

SALVUCCI, D. D. AND GOLDBERG, J. H. 2000. Identifying fixations and saccades in Eye-Tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM Press, 71–78.

STARKER, I. AND BOLT, R. A. 1990. A gaze-responsive self-disclosing display. In *CHI '90: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA, 3–10.

STELLMACH, S., NACKE, L., AND DACHSELT, R. 2010. 3D attentional maps: aggregated gaze visualizations in three-dimensional virtual environments. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI 2010)*. ACM Press, 345.

SUNDSTEDT, V., BERNHARD, M., STAVRAKIS, E., REINHARD, E., AND WIMMER, M. 2013. Visual attention and gaze behavior in games: An object-based approach. In *Game Analytics*, M. Seif El-Nasr, A. Drachen, and A. Canossa, Eds. Springer London, London, 543–583.

SUNDSTEDT, V., STAVRAKIS, E., WIMMER, M., AND REINHARD, E. 2008. A psychophysical study of fixation behavior in a computer game. In *APGV'08: Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*. ACM, 43–50.

TOLE, J. AND YOUNG, L. 1981. *Digital Filters for Saccade and Fixation Detection.* Lawrence Erlbaum, Hillsdale, NJ, 185–199.

TOMASI, C. AND MANDUCHI, R. 1998. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV'98)*. IEEE Computer Society, 839–846.

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

ŠPAKOV, O. 2011. Comparison of gaze-to-objects mapping algorithms. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications*. NGCA'11. ACM, New York, NY, USA, 6:1–6:8.

WEGHORST, H., HOOPER, G., AND GREENBERG, D. P. 1984. Improved computational methods for ray tracing. *ACM Transactions on Graphics 3,* 1, 52–69.

WOLFE, J. 2000. Visual attention. In *De Valois KK, editor. Seeing* 2nd Ed. Academic Press, San Diego, CA, 335–386.

XU, S., JIANG, H., AND LAU, F. C. 2008. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM conference on Recommender systems*. RecSys '08. ACM, New York, NY, USA, 83–90.

YANULEVSKAYA, V., UIJLINGS, J. R. R., GEUSEBROEK, J. M., SEBE, N., AND SMEULDERS, A. W. M. 2013. A proto-object-based computational model for visual saliency. *Journal of Vision 13,* 3.

ZHANG, X., REN, X., AND ZHA, H. 2008. Improving eye cursor's stability for eye pointing tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. ACM, New York, NY, USA, 525–534.