Reactive Gaze during Locomotion in Natural Environments

Julia K. Melgare, Damien Rohmer, Soraia Musse, Marie-Paule Cani

ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2024
M. Skouras and H. Wang
(Guest Editors)

Published at COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 8

Reactive Gaze during Locomotion in Natural Environments

J. K. Melgare[1,2], D. Rohmer[2], S. R. Musse[1] and M-P. Cani [2]

[1]PUCRS, School of Technology, Brazil
[2]École Polytechnique, CNRS (LIX), IP Paris, France

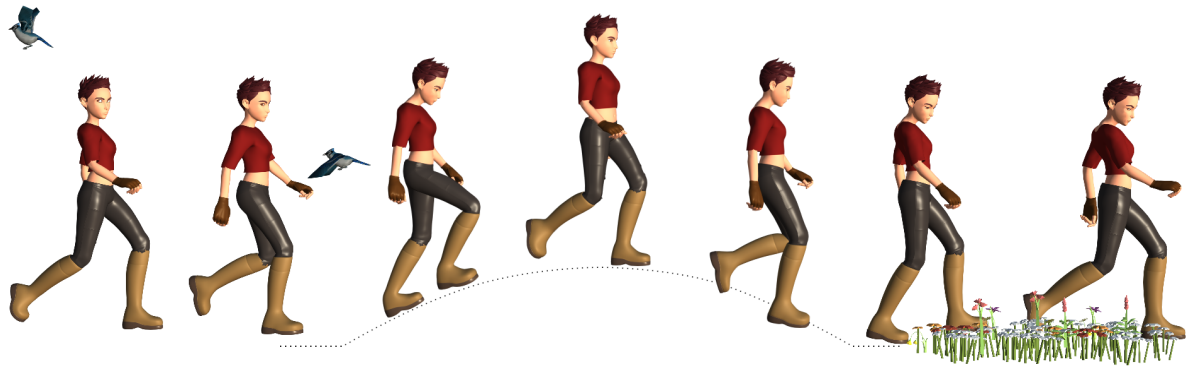

Figure 1: *Our method automatically generates a virtual character's gaze animation (eye and head movement), making it look aware of its environment. Our visual attention model considers the character's path, the slope of the terrain, and the salience of shapes and movements of surrounding elements.*

Abstract

Animating gaze behavior is crucial for creating believable virtual characters, providing insights into their perception and interaction with the environment. In this paper, we present an efficient yet natural-looking gaze animation model applicable to real-time walking characters exploring natural environments. We address the challenge of dynamic gaze adaptation by combining findings from neuroscience with a data-driven saliency model. Specifically, our model determines gaze focus by considering the character's locomotion, environment stimuli, and terrain conditions. Our model is compatible with both automatic navigation through pre-defined character trajectories and user-guided interactive locomotion, and can be configured according to the desired degree of visual exploration of the environment. Our perceptual evaluation shows that our solution significantly improves the state-of-the-art saliency-based gaze animation with respect to the character's apparent awareness of the environment, the naturalness of the motion, and the elements to which it pays attention.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics;*** *Animation;* ***Procedural animation;***

1. Introduction

Plausible gaze behavior, including eye and head motion, is crucial for creating believable animated virtual characters, providing clues about the character's perception and awareness of its environment [Bad97, CAC*22]. While visual attention and gaze animation have been extensively explored for conversational scenarios where virtual agents have to engage with the user or interact with each other [RPA*15, DOA22], very few work have focused on simulating visual attention for individual characters walking in an open space environment. Still, these scenarios are critical in video game applications and real-time simulations, where characters exploring and engaging in a dynamic environment are commonly met. The absence of an effective and general approach to manage gaze animation often leads to static heads and eyes staring blankly ahead. Thus, the characters do not seem to pay attention to their surroundings. In particular, this static behavior critically lacks realism when moving objects are present in the environment.

In this work, we aim to provide a natural-looking gaze animation model, able to simulate a dynamic character's visual attention and to automatically adapt its gaze behavior to the environment,

Authors' version of the article published at Computer Graphics Forum, vol. 43, n.8, Proceedings of the Symposium on Computer Animation (SCA), 2014.

in the context of real-time locomotion and exploration-looking behavior. We specifically consider the case of an open natural environment that highlights such conditions and provides a variety of visual stimuli ranging from colorful vegetation to moving animals, with varying terrain slopes requiring the character's attention. To remain useful in contexts such as games or VR applications, our approach is compatible with both automatic navigation following a pre-defined path and interactive locomotion guided by a user.

Handling gaze animation in such a context is challenging for the following reasons. On the one hand, neuroscience studies can describe human gaze behavior rules in specific test-case scenarios, including locomotion [TGCL20]; however, generalizing the latter to arbitrary open worlds where competing stimuli are acting is impractical as it would require defining a prohibitive number of dedicated rules to react to each possible element in the environment. On the other hand, data-driven saliency models can accurately predict visual attention related to shapes and colors [KSDG20] but fail to capture cognitive aspects of egocentric motion perception that need to take into account both peripheral vision and proprioceptive awareness of the viewer's own motion [KB23].

To address these limitations, we propose a holistic approach, coupling behavioral models for high-level cognitive decision and motion perception using a generic image-based saliency representation of the environment. We relied on the analysis of neuroscience literature to extract relevant individual behavioral models and parameters. We then built, for the first time, a combination of such behaviors within a single unified representation able to adapt the character's gaze to (a) its locomotion, integrating its curved path and the terrain slope condition, (b) the environment stimuli regarding both image-based and motion-based saliency. In addition, the sensibility to both of these states can be parameterized at a high level, enabling our method to seamlessly transition between characters carefully exploring a new environment and paying high attention to every detail, up to those traveling a familiar route and mostly focusing on their path.

Our technical contributions are twofold:
(i) An effective method, leveraging studies in motion perception, for estimating the saliency of moving objects (Sec. 4);
(ii) An attention decision system built principally from drift diffusion models (DDM) in neuroscience [RM08] to determine whether the walking agent should focus on an external object or on its path, based on current environmental conditions and its internal memory (Sec. 5);

After applying our model to various natural environments and scenarios, we validate the benefits of our approach via a user study (Sec. 6). This confirms that the generated gaze animation, improves the feeling of the character's awareness to its environment compared to the state of the art.

2. Related Work

Gaze animation requires two elements. First, detecting the visible environment using synthetic vision, and second, parsing it into a set of salient objects using *visual attention models*. In this section, we provide a concise overview of previous research on these two topics. We refer the reader to the surveys of Peters et al. [PCR*11] and Ruhland et al. [RPA*15] for a more complete review of the literature concerning synthetic vision and gaze animation.

2.1. Synthetic Vision for Virtual Agents

Renault et al. [RTT90] were the first to propose an approach for executing behavioral animation for agents based on synthetic vision. They referred to this approach as $2\frac{3}{4}$D vision, as it was produced by combining a representation of the scene through the agent's point of view in 2D and augmenting it with geometric information from the environment, such as an object's distance to the agent and its respective identifier. Later, other work started relying solely on image data to determine the agent's information about the scene to make synthetic vision more generalizable and realistic [CKB99].

Given the different strategies applied for modeling vision in virtual environments, Peters et al. [PCR*11] classified work about perception and visual attention for virtual agents into two main approaches. First, geometry-based approaches grant the agent direct geometric information about the scene, where the visibility is typically computed via ray casts. This approach may provide very accurate information compared to real limited human vision, and specific mechanisms have been developed to limit the agent's omniscience [YLNP12, AG18, EHSN19].

Second, image-based synthetic vision approaches consist of rendering the scene from the agent's point of view. Closer to human vision, this robust method can adapt to any type of environment. However, the associated low-level data (pixel colors) may be more difficult and expensive to analyze than a high-level description of the 3D scene [IDP06, NCRP16].

In our work, we use two types of vision models to balance adaptability and efficiency. An image-based representation computes visual attention on a static view, while the 3D scene description is used to efficiently analyze the movement of dynamic shapes.

2.2. Visual Attention and Gaze Animation

Visual attention is the cognitive process that mediates the selection of important information from the environment [LM21]. For virtual agents, both this process and the gaze animation that help communicating it can be either procedurally modeled based on behavioral studies from neuroscience or psychology [IDP06, GD02], or learned from eye-tracking data of real humans [ZZWT24].

Behavioral methods take inspiration from neuropsychology studies to create rule-based computational models for the agent's actions and behaviors. Itti et al. [IDP06]'s neurobiological visual attention model extracts color, light intensity, and motion from an input video stream to produce a saliency map for each video frame, used to generate gaze fixations for a virtual character. Their approach is simple and produces visually realistic results, but is not real-time so the animation needs to be preprocessed. Improving upon Khullar and Badler [CKB99]'s model, Gillies and Dodgson [GD02] propose a general parameterized model for visual attention capable of producing gaze animations for different scenarios, including navigation and locomotion. They use an attention manager, which receives attention requests from scene queries based on the agent's geometric vision and treats them to determine

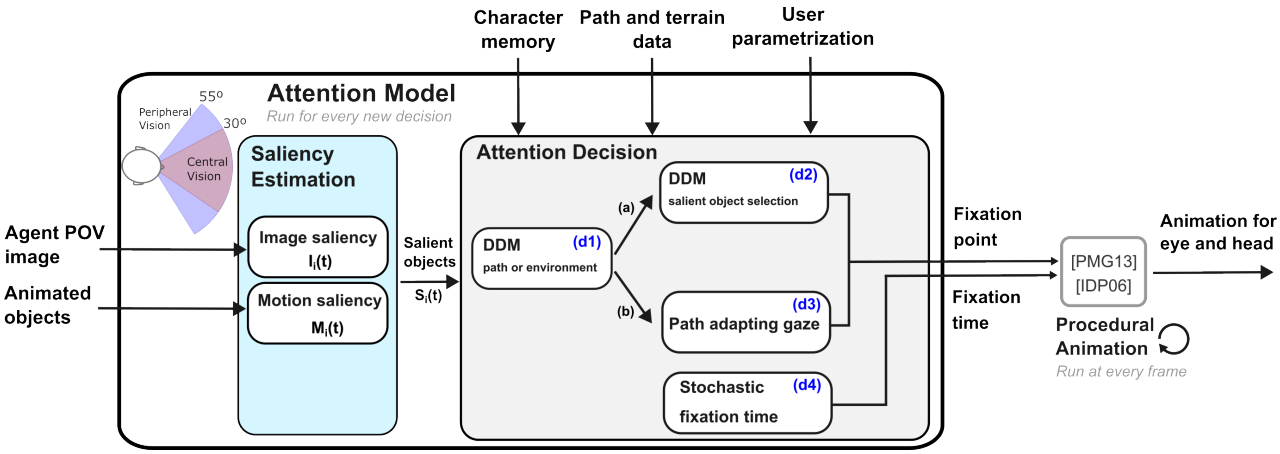

Figure 2: *Overview of the method. The attention model first relies on two saliency estimation modules, respectively focused on saliency in a static image and motion saliency. Given that the agent is waking, this allows taking a decision between looking at a specific salient object (a) or at the path (b), featuring the generic principle of Drift Diffusion Model (DDM). From this choice, the object or point of interest is selected and completed by a stochastic model of fixation time, which finally drives a procedural animation of the eyes and head.*

the focus of the agent's gaze. Although this model is easy to generalize, it requires heavy manual adjustments from the user, who is to provide both semantic information about the environment and fine parameter tuning to achieve the desired result. Aside from rule-based methods, other behavioral approaches also make use of traditional reinforcement learning techniques to model intrinsic cognitive processes behind the human gaze behavior [HB05]. Sprague et al. proposed an abstract behavioral model for embodying visuomotor behaviors of humans, particularly when executing navigational tasks that involve obstacle avoidance while walking [SBR07]. They propose an operating system that manages many learned microbehaviors that are applied in different abstraction levels to control a virtual human character.

Data-driven methods use neural networks trained on large datasets of images (or videos) annotated with human gaze or attention data, to estimate a saliency score representing the visual *attractiveness* of each pixel. We refer the reader to the MIT/Tübingen Saliency Benchmark [KBJ*] for an extensive study and benchmark of methods in this very active research topic. Close to our goals, Goudé et al. [GBO*23] applied Kroner et al. [KSDG20]'s visual saliency model to simulate visual attention and produce realistic gaze behavior for static, non-conversational agents. They combine the saliency map from a data-driven model with a saccadic map based on users' eye-tracking data to generate fixation points for the agent's gaze animation system. Although their results proved indistinguishable from performance captures, the method does not account for motion and cannot be applied to a walking agent.

Beyond the image saliency map, scan-paths estimations provide a time sequence of successive gaze fixation points [MSB*22] while video saliency generalize image-saliency to short video sequences, which can be achieved at interactive rates [WSX*19, ZZWT24]. Unfortunately the latter cannot be directly applied to egocentric walking character motion estimation, for two reasons: First, a video recording of a walker's view would capture not only the movement of external elements, but also the relative movement of the character's head in that environment, which would lead to the detection of spurious, non-salient movement for the human brain. Second, human vision relies on two components [SRJ11]: *Central vision* where all stimuli, including shape, color and movement, are visible, and *Peripheral vision*, giving more limited access to shapes and colors, but remaining very sensitive to movement. As video saliency methods rely on a coupling between image color and motion, they cannot faithfully model peripheral vision mechanisms. To avoid this limitation, we combine the two saliency estimation strategies, using a data-based model to analyze shape and color in central vision and a geometry-based model to calculate motion saliency, including in areas of peripheral vision.

3. Method overview

The input of our method is an animated character walking in a virtual environment. Its trajectory may be either pre-set, or interactively controlled. Our processing pipeline for animating the head and gaze is summarized in Fig. 2. Given the current head position, we run our *attention model* every time the character needs to choose a new focus of attention. Our model includes two main modules:

First, the *saliency estimation* module evaluates the perception of all visual elements from the walking character's point of view while remaining agnostic to high-level character memory and locomotion. This includes both static saliency, computed on an image restricted to the central vision zone, and saliency due to motion, computed in both central and peripheral vision zones (see Section 4).

From this estimation, we generate a list of objects with associated saliency scores that are potentially attractive to the character's gaze.

Second, the *Attention Decision* process handles competing stimuli such as salient features of the environment, the difficulty of the terrain the character is walking on, and the character's internal state and memory (see Sec. 5). The attention decision model calculates

a fixation point, i.e. a specific target point in the 3D scene that becomes the focal point of gaze, as well as a fixation time associated to this point. The decision about the fixation point is itself composed of three sub-steps. First, a binary decision (d1) selects whether the character should look at a salient object of the environment (a) or at the path (b). In the first case, a second decision process (d2) is initiated to select which specific object of the environment will be targeted by the character's gaze while taking its short-term visual memory into account through the notion of *Inhibition Of Return*. Conversely, if choice (b) is selected, a procedural approach (d3) is used to define the exact point on the path that the character will focus on. The resulting fixation point is associated with a fixation time calculated using a stochastic approach (d4).

Once a fixation point and time are set, existing methods are combined to generate the final animation towards that point, including coordinated head and eye movements [PMG13] as well as eyes blinking [IDP06].

4. Motion Aware Saliency Estimation

The first component of our model is the *saliency estimation* module, which computes the character's ability to perceive its environment via a synthetic vision system associated with saliency score evaluation. For a time t and an object of the environment designated by its index i, we call $S_i(t) \geq 0$ its saliency score (a positive scalar value). The calculation of $S_i(t)$ is shared between two specialized estimators: the first one $I_i(t)$, dedicated to static saliency in a single image, relies on an existing data-driven approach; the second one $M_i(t)$ is a new estimation of the saliency of moving object, introduced in this work. The output of saliency estimation at a time t is a list of N_{salient} objects with their associated scores $S_i(t)$ defined as

$$\forall i \in [\![1, N_{\text{salient}}]\!],\ S_i(t) = I_i(t) + M_i(t). \tag{1}$$

In the following, we detail the computation of these two saliency scores and their associated synthetic vision models.

4.1. Static image-based saliency $I_i(t)$

The saliency related to shape and color of static elements of the scene is computed using an image-based representation of the central vision. To this end, we use a rendered view of the environment from the current frame centered at the character's mid-eye with a field of view of 60 degrees, which corresponds to the typical range of the central vision in humans [SWHH90]. This region is considered to be the central vision of the agent. We then use the state-of-art method from Kroner et al. [KSDG20] featuring an encoder-decoder neural network able to infer a saliency map on top of the rendered image at interactive rates. The neural network is pre-trained on real images, which deviates slightly from the cartoon-style rendering used in our application. Nevertheless, employing this network in virtual environments was previously documented in the literature and shown to provide accurate results [GBO*23, MMA*23]. Once the saliency map is computed, we consider regions with high values, i.e., above 50% relative to the maximal saliency value over the whole scene, and retrieve the associated object from the scene via ray-casting. Each subsequently selected object defined by its index i is associated with a saliency value $I_i(t) \geq 0$, corresponding to the maximum saliency of the pixels associated to that object in the saliency map.

4.2. Motion saliency for animated elements $M_i(t)$

Complementary to static, image-based saliency, we propose a per-object motion saliency characterized by the value $M_i(t) \geq 0$ computed over both central and peripheral vision. The latter is approximated by a conical frustum with a field of view of 110 degrees, given the range of the visual field of real humans [SWHH90]. As mentioned in Section 2, video-based models cannot capture motion-only saliency. Instead, we propose a fast geometric approach allowing to account for all the dynamic elements of the 3D scene whose bounding box intersects the cone of peripheral vision.

Our model uses the characterization proposed in the perceptual study of Arpa et al. [ABC11] defining five different types of perceived motions:

- *Appearance*: The object appears on the screen for the first time.
- *Onset*: The start of motion (transition from static to dynamic)
- *Change*: The change of the object's speed or direction;
- *Offset*: The end of motion (transition from dynamic to static)
- *Continuous*: The object is moving with the same velocity;

The three first states are perceived as the most salient, whereas the two last are perceived as less salient, with a quantitative ratio of 5:1. This characterization is useful as it provides a generic way to interpret motion with respect to its perceived saliency. Notably, human perception of motion is more influenced by the type and changes in motion rather than the absolute speed of the movement. For instance, an element moving rapidly in a straight line tends to attract less attention than a sudden change of trajectory (or the start of a motion) of another slower element. Still, this study and their associated saliency score were limited to rigid spheres moving at a constant distance from the viewer [ABC11]. Consequently, they do not fully account for the complex motions seen in natural environments, such as animals with moving parts that can be at different distances from the character. To address this, we propose to extend the characterization of such perceptual motion to include rigged objects where the motion is linked to an animated skeleton.

Let us consider a typical animation skeleton composed of a set of N_{joint} joints. The root joint, assumed to be at index 0, is associated with an absolute translation in the world space, while the child's joints at indices $j > 0$ are associated with rotations. Let us call $\vec{v_0}(t) \in \mathbb{R}^3$ the linear velocity vector of the root joint at time t. Seeing the skeleton as a kinematic chain, the world-space velocity of a joint j can be expressed as

$$\forall j \in [\![0, N_{\text{joint}}]\!],\ \vec{v_j}(t) = \vec{v_0}(t) + \sum_{k \in \mathcal{A}(j)} \vec{L_k} \times \vec{\Omega_k}(t), \tag{2}$$

where $\mathcal{A}(j)$ is the list of ancestor joints of j up to the root, L_k is the bone vector with extremity k, and $\vec{\Omega_k}$ is the angular velocity of joint k.

This world space velocity is then converted to an *apparent velocity* $v_j^a(t)$ from the character's camera, considering the effect of perspective, assumed to vary as $1/d_j(t)$, where $d_j(t)$ is the distance between the character's eyes and the joint j:

$$\vec{v}^a_j(t) = \vec{v}_j(t)/d_j(t). \tag{3}$$

We further associate the apparent acceleration of each joint defined with the finite difference $\vec{a}^a_j(t) = (\vec{v}^a_j(t) - \vec{v}^a_j(t-\Delta t))/\Delta t$.

We then compute a per-joint perceptual motion-saliency coefficient $m_j(t)$ following the perceptual rules introduced previously, and associated with the limit threshold $\epsilon_v > 0$ (resp. $\epsilon_a > 0$) from which a specific velocity (resp. acceleration) is perceived

$$m_j(t) = \begin{cases} 1 & \text{if } \|\vec{a}^a_j(t)\| \geq \epsilon_a \;\&\; \|\vec{v}^a_j(t-\Delta t)\| \geq \epsilon_v, \\ 1 & \text{if } \|\vec{a}^a_j(t)\| \geq \epsilon_a \;\&\; \|\vec{v}^a_j(t-\Delta t)\| \leq \epsilon_v, \\ 0.2 & \text{if } \|\vec{a}^a_j(t)\| \leq \epsilon_v \;\&\; \|\vec{v}^a_j(t-\Delta t)\| \geq \epsilon_v. \end{cases} \tag{4}$$

The first condition corresponds to the motion type *Change*, the second one to *Onset* or *Appearance*, and the third one to *Offset* and *Continuous*. We define the notion of *perceptual speed* of a joint $V_j(t) > 0$ as the product between the average speed of the joint and its perceptual motion-saliency coefficient.

$$V_j(t) = m_j(t)\,(\|\vec{v}^a_j(t)\| + \|\vec{v}^a_j(t-\Delta t)\|)/2\,. \tag{5}$$

We finally define the salience of an entire object as that of its joint of maximum *perceptual speed*:

$$\begin{cases} M_i(t) = m_{j_{\max}}(t) \\ \text{where } j_{\max} = \underset{j \in [\![0, N_{joint}-1]\!]}{\operatorname{argmax}} V_j(t) \end{cases} \tag{6}$$

5. Attention Decision Model

We now describe the second part of our solution, modeling how a visual decision is made about where to look, based on different stimuli such as the previously calculated saliency, but also the current locomotion of the character, the difficulty of the terrain and the current state of the character's visual memory. This decision system aims to emulate natural human-like behavior. We therefore integrate a certain degree of randomness, while providing a coherent response to various stimuli that may correspond to competing objectives.

5.1. Drift Diffusion Model for Generic Cognitive Decision

We propose to leverage the so-called *Drift Diffusion Model* (DDM), a well-established tool in cognitive sciences to explain the temporal processes linked to decision-making [Rat78, RM08]. DDM is based on the assumption that evidence in favor of each possible choice accumulates over time (drift). Each drift process is perturbed by random fluctuations following a normal law (diffusion), and a decision is made as soon as enough evidence supporting an alternative has accumulated, which simulates human decision-making.

For the sake of completeness, we first summarize the general framework of a multi-objective DDM process before detailing how we adapt this formulation for the character's gaze decision. Let us assume a set of N_{goal} possible scalar goal-related scores $(g_i)_{i \in [\![1, N_{\text{goal}}]\!]}$ where each g_i is initialized to 0 at the beginning of the decision process, and varies in $[-\lambda, \lambda]$ during the decision iterations, where λ is a scalar decision threshold value which depends on the DDM process. At each time step, between t to $t + \Delta t$, the goal-related score is updated from its current state as defined in Equation 7:

$$\forall i \in [\![1, N_{\text{goal}}]\!],\; g_i(t+\Delta t) = g_i(t) + (\mu_i(t) + r\mathcal{R} + b)\,\Delta t, \tag{7}$$

where μ_i is a *drift process* that depends on the stimuli influencing the decision process and corresponds to a score of attraction toward the goal g_i. $\mathcal{R}$ is a random process following a unit normal distribution, while r indicates the magnitude of the random fluctuation. b is an optional bias that can model an internal constraint of the decision process beyond the one conveyed by the external stimuli. At run time, each individual score g_i from all possible objectives are updated in parallel, and the first to reach the decision threshold $g_i = \lambda$ is considered to be the next ongoing decision. Then, all scores are re-initialized, and a new decision process can occur.

5.2. Drift Diffusion Model for Gaze Fixation

To apply the DDM to gaze fixation, we first define the notion of *inhibited saliency* $\hat{S}_i(t)$. While the saliency estimator previously presented identifies a set of interesting objects based on synthetic vision, their saliency scores are independent from the cognitive aspects of the character's memory. To take this into account, we consider the so-called *Inhibition of Return* (IOR), stating that points that have recently been the focus of attention are temporarily inhibited for a brief period of time [ABC11]. We then define the *inhibited saliency* $\hat{S}_i$ of the object i as

$$\hat{S}_i(t) = \omega_i(t)\, S_i(t), \tag{8}$$

where $\omega_i \in [0, 1]$ represent the IOR factor. We consider a typical inhibition time $\tau_{IOR} = 900ms$, and set $\omega_i(t) = \min(\delta t_i(t)/\tau_{IOR}, 1)$, where $\delta t_i(t)$ is the elapsed time since the object i was fixed.

As shown in Fig. 2, our decision mechanism is performed in two steps. First, we decide that the character should either look at the path or at a salient object from the environment, using a first DDM associated with a goal $g^{p/e}$ and a drift process $\mu^{p/e}$. If the decision relates to a salient object from the environment (choice (a)), we run a second multi-objective DDM over the N_{salient} objects with goal $g_i^{salient}$ and drift process $\mu_i^{salient}$ to select which salient object of the environment is the next target. Otherwise, if the decision relates to looking at the path, we define the fixation using procedural rules detailed in Sec. 5.3. A decision process is started every time a fixation period ends. Then, we allocate a 200ms time-window to make a new fixation decision, which aligns with the typical duration for the human eye to select a new target and shift towards it (see [PDGea01]). Only during this allocated time, the drift process and its respective goal-related scores are computed and updated every frame. Until a decision is made, the agent will remain focusing on the last fixation. If the decision process did not converge in the 200ms time-period, we use the object of current maximal score as the next focus of visual attention.

First DDM (d1 in Fig. 2): The choice between salient object or path takes into account two main notions mentioned in the neuroscience literature [HPV02, TGCL20]: The notion of terrain difficulty that we characterize as the local slope of the terrain at the character's position α, and the elapsed time since the character last

looked at the path δt_{path}. We then propose a single drift process, taking either positive or negative values, to link these notions with the inhibited saliency from Eq. 8:

$$\mu^{p/e}(t) = \alpha(t)\ \delta t_{\text{path}}(t) - \frac{1}{N_s}\sum_{i=0}^{N_{\text{salient}}} \hat{S}_i(t)\ . \tag{9}$$

The positive contribution $\alpha(t)\ \delta t_{\text{path}}(t)$ is computed using normalized values in the range $[0,1]$ and acts as an attraction toward looking at the path. The negative contribution depends on the averaged inhibited saliency and contributes toward looking at a salient object of the environment. The decision is taken based on the associated goal score $g^{p/e}$ corresponding to a binary goal DDM. If $g^{p/e}$ reaches this threshold value $\lambda^{p/e} = 0.5$ [Rat78] (or a positive value after 200ms) the winning decision is to look at the path, while conversely, if $g^{p/e}$ reaches the value $-\lambda^{p/e}$ (or a negative value after 200ms), the winning decision is to look at a salient object.

Additionally, we set the bias b defined in Eq. 7 to be a positive or negative value in order to favor more or less the decision to look at the path or at the salient objects in the environment. This bias enables us to parameterize, at a high level, the typical behavior exhibited by the character. With a positive bias, the character would tend to look more at the path, which we refer to as a Goal-Driven behaviour. In contrast, a negative bias makes the character more likely to look at external stimuli, and we call this behavior Exploratory. The effect and differences between these behaviors are analyzed in Section 6.2.2.

Second DDM (d2 in Fig. 2): Selecting a fixation point in the environment is obtained using a drift process taking directly into account the inhibited saliency and set as:

$$\mu_i^{salient}(t) = \hat{S}_i(t). \tag{10}$$

The most salient object is identified by the index i_0 in the multi-objective DDM such that $g_{i_0}^{salient}$ is the first to reach the threshold value $\lambda^{salient} = 0.2$ [RRHO23] before 200ms, or associated with the largest value after this time. In this decision process, the bias is set to 0, and the 3D point corresponding to the gaze fixation is set to be at the collision point through which the object was detected during the vision ray-casting.

5.3. Procedural Path-Adapting Gaze Fixation Point (d3)

When the character decides to look at its path, we rely on a procedural approach to compute the point of focus, grounded from the following key observations: First, the gaze direction anticipates the head orientation, which itself, anticipates the orientation of the main body when following a curved trajectory [BKB*12]. Second, on a flat terrain, our eyes are typically looking at about 7 to 8 steps ahead, while we shift down our focus to about 2 or 3 steps ahead when the terrain becomes uneven [TGCL20]. Third, a study performed on stairs [MdAM11] (assimilated to a slope at 30°) showed that we look 2 to 4 steps ahead when going up and down.

Based on these observations, we propose a model where the character anticipates its path by looking at a distance d ahead of its expected trajectory, with d depending linearly on the slope of the terrain. In our case, we consider

$$d = d_{\max} - \frac{\alpha}{\alpha_{\max}}(d_{\max} - d_{\min}), \tag{11}$$

where α is the current slope of the terrain at the character position, $\alpha_{\max}$ is the maximal slope considered as 30°, and $(d_{\max}, d_{\min})$ are respectively (7 steps, 3 steps) as reported in the literature, and can be converted to actual distance in meter depending on the length of the character's legs.

This model can then be used in two different contexts. First, if the character's trajectory is already fully preset, we simply query the associated point along the trajectory at the specific prescribed distance. Second, if the trajectory is dynamically controlled by a user – such as in game-like control – we infer an estimated spline trajectory based on current orientation and user input, which is then used to compute the future estimated point of focus.

5.4. Fixation Duration (d4)

In parallel with the computation of the fixation point, the time the agent will look at it is computed via a stochastic approach. Following the cognitive study from Droll et al. [DE09], we consider a different normal distribution of fixation time depending on the value of the bias b. In the case of a goal-driven behavior (positive value of b), we consider a distribution centered at 121ms and standard deviation of 67ms. In the opposite case, where the character is explorative (negative value of b), we use a distribution with an average of 1869ms and a standard deviation of 998ms. In the default, non-biased mode ($b = 0$), we randomly select one of these distributions to sample from.

Finally, and to account for the differences in anticipation time between the eye and the head observed in the literature [BKB*12], we introduce a delay in the animation model, where the head only starts to move towards the path fixation target at a normal random interval between 200 and 300ms after the eyes have started to move. This additional delay is only introduced when the agent is fixating on the path. For other fixation targets, we use the head-eye coordination delay established in the literature [PMG13].

6. Results and Evaluation

In this section, we show some of our results obtained in dynamic natural environments featuring moving animals and salient vegetation. We also present the perceptual evaluations conducted to assess whether our model improves upon the current state of the art for saliency-driven gaze animation and if users are able to differentiate between Exploratory and Goal-Driven behaviors.

6.1. Implementation and featured scenarios

We implemented our animation and decision process in Unity 3D and coupled it with the image-based saliency model run in Python/TensorFlow at interactive time in parallel with the animation. Our model is used to compute the eye and head animation, while the rest of the body, i.e. torso, legs and arms were animated using an existing approach [ARC22]. We considered the case of two main scenes that featured rigged-animated animals such as

birds and butterflies, colorful vegetation such as mushrooms and flowers, and uneven terrain heights. Our interactive demo, which allows for user manipulation of the character through a keyboard or game-pad, runs in real-time on a standard laptop (Intel Core i7, eight cores, running at 3.10 GHz). Table 1 presents the breakdown of our computational cost and refresh rate for the different parts of our algorithm. These timings are reported as an average over 1500 frames. Steps 1 to 4 are computed every 200ms, to simulate the rate at which humans shift between saccades, as mentioned in Section 5. Step 5 starts whenever a new fixation decision needs to be determined, and the DDMs are computed every frame while the process of decision-taking is being simulated (with a maximum duration of 200ms). Finally, the facial animation routine is computed every frame. Most of our decision process have run time below 0.1ms, and the pre-trained image saliency evaluation takes about 0.2ms while being run every 200ms. The highest computational cost (5ms) currently relates to the image data transfert from the character point of view between Unity and Python on which the image saliency is run as an external program. This transfert is currently done trivially but could be heavily optimized via shared memory use if the method need to be scaled to multiple characters.

Step	*Avg Computation Time (ms)*	*Evaluation Frequency*
1. Motion Saliency ($M_i(t)$)	0.1 (SD = 0.07)	Once every 200ms
2. Data transfer Unity/Python	4.91 (SD = 0.58)	Once every 200ms
3. Image Saliency ($I_i(t)$)	0.22 (SD = 0.28)	Once every 200ms
4. Attention Decision	0.07 (SD = 0.07)	Whenever a new fixation is needed
5. Facial Animation	0.02 (SD = 0.1)	Every frame

Table 1: *Performance analysis of our method.*

Figure 1 depicts the main behavior of our character able to, respectively, look at the flying bird with motion saliency, check more carefully her path on uneven terrain, and pay attention to salient elements such as colorful flowers. We show our full test scenes in Figure 3 using a neutral value for the bias b. Scenes (a) and (b) featured 4 to 5 animated flying animals (birds and butterflies), each with their own behaviors (i.e. butterflies flying close to the flowers, birds flying away from the agent when approached). The salient vegetation is spread throughout, and their terrain is uneven with slight elevations along the path. The terrain in scene (c) was made to be more uneven in order to focus on our path-adapting gaze, with a slope of 30 degrees.

Our method yields animations that depict the character as both engaged in its walking task and curious about exploring its surrounding environment. The addition of motion perception makes the agent seem much more attentive as it is capable of reacting to fast movement and tracking moving objects while fixating on them, as seen in Figure 3(a) and (b). At the same time, the path anticipation and the adaptation of the gaze according to the current slope angle are subtle changes that contribute to make the animation more natural for a walking agent (Figure 3(c)).

In Figure 4, we demonstrate the different animation outcomes obtained when alternating between Exploratory (i.e. paying careful attention to salient element of the environment) and Goal-Driven behaviors (i.e. principally focus on its path) by altering the bias b on the choice between focusing on the path or on external objects. In the examples shown, we compare two simulations applied to the same scene, where the exploratory agent fixates on the butterflies as she walks, while the goal-driven agent focuses on the path and anticipates where to look within it. The changing of the fixation time distribution according to the set behavior also greatly affects the outcome animations. For instance, the exploratory agents lingers for longer amounts of time at each fixation, while the goal-driven agent only shoots quick glances to other objects before going back to focusing on the path. We further evaluate if these behaviors are differentiable to users in a perceptual study described in Section 6.2.2. Finally, Figure 5 shows an ablation study applied to the example of Figure 1. As can be seen, the ability to be drawn to movement and image-based salience as well as changes in terrain helps make the character appear more aware of their surroundings. Even though the character's body motion remains unaffected by our approach, having a responsive gaze animation alone already improves the expressiveness of the character, especially when compared to a static, forward-facing gaze animation.

6.2. Perceptual Evaluation

We conducted two perceptual evaluations to assess the following: *i*) whether or not our model can improve on the current state of the art when it comes to simulating gaze behavior during locomotion in dynamic environments, and *ii*) whether users are able to differentiate between our two main modeled behaviors (Goal-Driven and Exploratory). We conducted two separate online surveys for each evaluation, where users were shown video recordings of the same scenes with different animation configurations in random order. In both evaluations, we chose to use a stylized character appearance to avoid the uncanny valley effect [Mor70], with short hair to facilitate viewing of the eyes and facial features. All animation configurations used the same walking speed of 1.2 m/s, with the same neutral walking style.

6.2.1. Comparison with the State of the Art

Our first evaluation assesses the perceived **awareness** of the character [GPMJ13], the **naturalness** of gaze animation [BKCZ09] and the plausibility of the locations which the character looks at [LHOK18], which we refer to as **gaze target**. We compare our work with the current state of the art in real-time saliency-driven gaze animation described in Goudé et al. [GBO*23] aimed at generating realistic gaze animations for static agents in static environments. We formulated the following hypotheses:

H1. Our model provides a stronger feeling of awareness of the character to its environment and improves the plausibility of the focused gaze target;
H2. Our model seems, at least, as natural as the current state of the art.

While our method was designed to improve the criteria mentioned in H1, it was not specifically intended to increase the global naturalness of the character, as the current literature reported performance already comparable to those of real actors. Still, we aim to confirm through H2 that our approach does not degrade the overall plausibility of the character's animation.

To perform this evaluation, we chose to compare three different configurations: *i*) our complete gaze animation model; *ii*) the

(a) *The agent first looks at the path as they are crossing the bridge, then looks at the flying birds on the following frames.*

(b) *The agent looks at a protruding mushroom, the butterfly in motion, then resumes anticipation of its trajectory.*

(c) *Its gaze first drawn to the flowers, the agent focuses more and more on the steep terrain as the slope increases.*

Figure 3: *Some examples of the results obtained when applying our method to dynamic, natural environments using a neutral bias $b = 0$. We highlight the agent's current fixation point in each frame with the white disk.*

(a) *Examples of Exploratory Behavior*

(b) *Examples of Goal-Driven Behavior*

Figure 4: *Comparison between Exploratory (a) and Goal-Driven (b) behaviors. In (a), the agent pays attention to the butterflies as they walk, which are external stimuli. Meanwhile, in (b), the agent focuses on anticipating the path.*

current state of the art [GBO*23]; *iii*) a baseline configuration where the character always looked to the forward direction. All configurations used the same walking model for the body animation [ARC22] and followed the same pre-defined path trajectory. Our model was presented using the "Exploratory" behavior. We implemented the state of the art approach in Unity 3D and added an additional behavior that would have the agent re-orientate its head upwards towards a neutral position every few seconds to compensate for the fact that the model was not originally conceived to deal with walking agents and would often look at the ground due to it being detected as salient. We used the two scenes created for our demo in this study, which are shown in Figure 3 (a) and (b), as well as in the supplementary video. They present both static and dynamic salient elements throughout the character's path.

Task: First, the participants were shown a video of the scene without any walking character, and were told to observe it, noticing what elements seem to attract their attention. Then, they were shown three 15-second videos in a side-by-side layout, each showing a different configuration for the same scene. The order in which the videos were presented was randomized. Aside from the normal scene view, we also included a close-up view of the character's

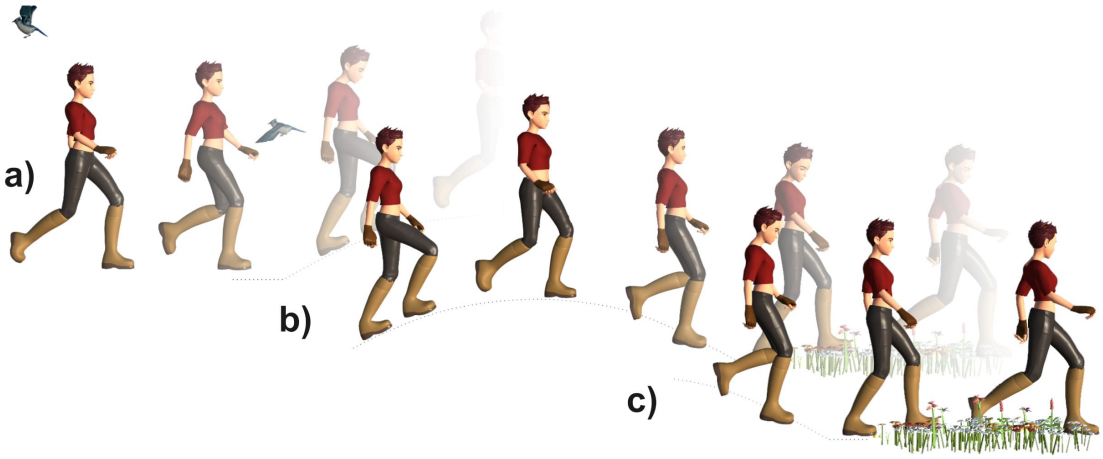

Figure 5: *Ablation study applied on the example from Fig. 1. a) No motion saliency: the character misses the flying bird. b) No dependence on terrain slope when looking at the path: the character looks straight on the bump. c) No image-based saliency: the character misses the flower on the ground.*

face and a highlight of the current fixation point to facilitate the participant's viewing task. The participants were allowed to pause and re-watch the videos as many times as they wanted throughout the experiment. For each configuration video (labeled as A, B and C), they were asked about how much they agreed with the three following statements (answered on a 5-point Likert-scale ranging from 1: strongly disagree to 5: strongly agree), designed to judge the animations given our criteria of awareness, gaze target and naturalness, respectively:

S1. The character seems aware of her environment
S2. The elements at which the character looked at made sense
S3. The movement of the character's eyes and head seems natural

Protocol: Upon starting the questionnaire, the participants were asked about their age, educational level, gender and familiarity with computer graphics (i.e. if they watch movies/videos with animated CG characters, play video games, etc.). Each participant was randomly assigned to a scene in order to respond to the statement questions. The scenes were evenly distributed among the participants.
Participants: Our questionnaire was distributed among colleagues and lay people through a custom online distribution system. After discarding incomplete answers, we had a final sample of 95 subjects (48 F, 45 M, 2 NB; 46% between 36 to 51 years old; 72% familiar with computer graphics; 52% post-graduates).

Analysis: Given our sample size and the nature of our data, we conducted Friedman tests to evaluate the effect of the different configurations on the perceived awareness, naturalness and on how much people agreed with the places where the character looked at. Post-hoc comparisons were performed using the Durbin-Conover test. The full details of our statistical analysis are available in the supplementary material.

Our Friedman tests indicated that a main effect was found among all evaluated aspects: Awareness ($\mathcal{X}^2(2) = 47.88, p < 0.001$), Naturalness ($\mathcal{X}^2(2) = 23.42, p < 0.001$) and Gaze Target ($\mathcal{X}^2(2) = 47.82, p < 0.001$). Our post-hoc analyses revealed that our method obtained a significantly higher score in Awareness (S1), Naturalness (S2) and Gaze Target (S3) when compared to both the state of the art and the baseline ($p < 0.001$). Figure 6 shows the score distribution for each configuration and evaluated aspect. These results show that our method was able to improve upon the state of the art in gaze animation in all of the evaluated aspects in the scenario of exploratory locomotion in a dynamic environment, which means that both of our formulated hypotheses (H1 and H2) were supported. Regarding H2, we initially did not anticipate our model to outperform the state of the art in terms of perceived naturalness. However, we believe that its adaptation to dynamic stimuli contributed to users perceiving the animation as more natural in the proposed scenarios.

6.2.2. Perception of Exploratory and Goal-Driven Behaviors

We conducted an additional perceptual evaluation in order to determine if users would be able to differentiate between the two extreme gaze behaviors that our model can display: Exploratory and Goal-Driven. Exploratory behavior means that the agent will prioritize looking at external stimuli away from the path, and will spend longer amounts of time looking at each object. In contrast, the Goal-Driven agent will prioritize looking at the path instead, and the fixation times when looking at external stimuli are shorter. Thus, our hypothesis for this perceptual study is that users will be able to correctly label a given behavior as either Exploratory or Goal-Driven when presented the respective stimuli.

Task: The participants were shown two 15-second videos in a side-by-side layout, each showing either the Exploratory or Goal-Driven behaviors. The layout followed the same pattern as in the previous study, with the difference being that we did not provide an additional close-up view of the character's face and did not highlight the current fixation point. The videos shown to the users in this experiment are also available in the supplementary video. Once again, the participants were allowed to re-watch the video as many times as they wanted throughout the experiment. For each of the

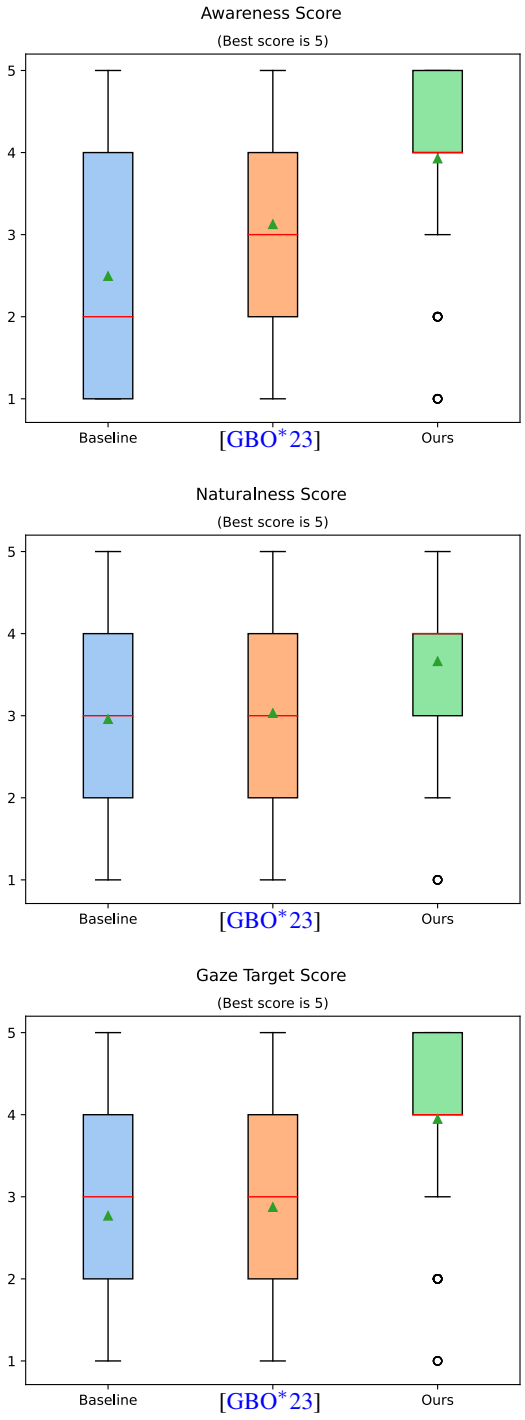

Figure 6: *Scoring of each tested configuration in our comparative user study. The median is represented by the red line and the mean is represented by the green triangle.*

two videos (labeled as A and B), the participants were asked to identify the character's demeanor as either "Exploratory", "Goal-Driven", or "I do not know". For contextualization, the following definitions were presented:

- Exploratory: "The character is exploring the path as if it was the first time she is seeing it."
- Goal-Driven: "The character is already familiar with the path and just wants to pass through it as she usually does."

Protocol: We used the same protocol as in the previous experiment. The response alternatives of either "Exploratory" or "Goal-Driven" were presented in random order to the participants, and "I do not know" was always presented as the last option.
Participants: Using the same online distribution system, we had a final sample of 310 participants (173 F, 134 M, 3 NB; 53% between 36 and 51 years old; 75% familiar with computer graphics; 47% post-graduates).

Analysis: After collecting all the participant responses, we obtained the results seen in Figure 7. As can be seen, the majority of the users were able to provide the correct response for both Goal-Driven and Exploratory videos. Chi-squared analyses indicated a significant difference in the response distribution for both behaviors (Goal-driven: $\mathcal{X}^2 = 281.46$, $p < 0.001$; Exploratory: $\mathcal{X}^2 = 282.77$, $p < 0.001$). These results show that, when shown the two animations with different behaviors, users were able to easily identify them with the proper label, which supports our hypothesis for this additional user study. The fact that a simple change in one parameter of our model is capable of producing two easily differentiable behaviors in the eyes of users means that our model can be applied in different animation scenarios and adapt to the given context. For instance, if animators or game designers wanted to have groups of characters walking in the same environment, they would be able to easily configure characters that would seem more busy or focused (Goal-Driven), versus more curious or distracted (Exploratory), through the same animation model.

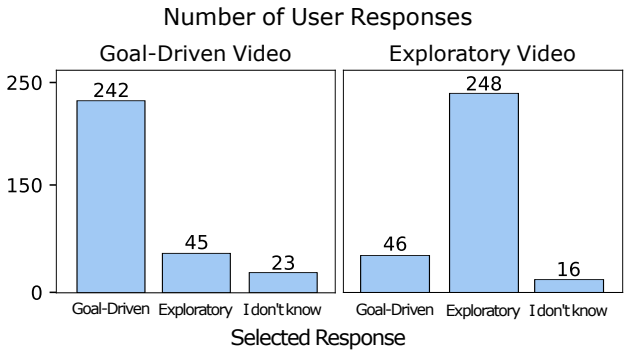

Figure 7: *Collected responses for the second user study. The graphs show the number of responses for both Exploratory and Goal-Driven videos.*

7. Discussion and Final Remarks

In this paper, we introduced a real-time visual attention model adapted for walking characters in dynamic virtual environments. Our approach integrates motion-aware saliency estimation with a decision-making process that combines the Drift Diffusion Model and procedural techniques based on neuroscience principles. We achieve gaze animations that convey a character's awareness of its surroundings and engagement with its locomotion task. Through perceptual evaluations, we demonstrated the efficacy of our approach in enhancing the believability of virtual characters. Our animations exhibit improved awareness of the environment, naturalness in motion, and plausibility of the chosen gaze targets compared to the state of the art in a context of locomotion in natural dynamic environments. We have also shown that the apparent behavior of the character can be influenced towards either exploration-focused or path-focused by altering one single parameter of the model, and such a change is shown to be easily perceptible by users.

Our contribution to the estimation of motion saliency of animated objects relied on a geometric approach. Although this constitutes a simple and generalizable solution for dynamic scenes, video processing via the use of optical flow would enable to take into account the loss of precision of peripheral vision, and could therefore more precisely simulate cognitive processing. A flaw of our current solution is that it does not measure contrast between motions: While our motion saliency model would work in most virtual scenes used in games and simulations, where the environment is mostly static, it could not be applied, for instance, if the character was inside a moving vehicle, where all the object do move but at the same speed, and therefore none are salient. Our second contribution, focusing on the decision-making process of the agent, relies on DDMs to simulate a variable behavior that is still capable of taking clues from its environment and internal state to choose between multiple plausible fixation targets. We relied on neuroscience literature to integrate the main factors influencing the associated drift process. While our current prototype only integrates the notion of slope and curvature of the terrain, our method could easily be extended by adding additional parameters to the drift-diffusion model, which is, by nature, able to handle multiple competitive senses to take a global decision. For instance, walking on difficult terrain could be integrated via dedicated gaze attention on the ground near the feet (e.g. stepping stones, river crossing, muddy or snowy terrain). Additionally, internal parameters of the character, such as its speed, could further influence the decision process.

In this study, we focused on applying our model to natural environments, as they allowed us to have a rich variety of stimuli for our agent without the need to rely on social rules that often influence our behavior. However, our model could still be extended to other environments given adaptations to the saliency estimation or to the drift processes. For example, a locomotion context in an urban environment where other agents may also be present could be modeled by including known concepts of human group or crowd behavior in the DDMs. Additionally, the image-based saliency model could be re-trained to better adapt to the scenario, where stimuli such as street signs and other people are more salient due to contextual cues.

For future developments, we envision the integration of our visual attention model with reinforcement learning techniques. This integration would enable an agent to learn to navigate and explore the environment using the information provided by our system. Such an approach would also permit the system to guide the character's movement based on visual inputs. Additionally, coupling gaze animation with a full-body locomotion model would enable to improve the way the character's shoulders and torso adjust when walking, observing interesting objects, or anticipating their path.

8. Acknowledgements

This project was funded by the CAPES Institutional Program of Internationalization (PrInt) grant (nº 88887.840444/2023-00) and CNPq. An exchange period was partly funded by Marie-Paule Cani's fellowship on Creative AI from Hi!Paris. We would like to thank Polyana Graf Finamor Correia for modeling and rigging the 3D virtual character used in this work.

References

[ABC11] ARPA S., BULBUL A., CAPIN T.: A decision theoretic approach to motion saliency in computer animations. In *Motion in Games* (2011), pp. 168–179. 4, 5

[AG18] AĞIL U., GÜDÜKBAY U.: A group-based approach for gaze behavior of virtual crowds incorporating personalities. *Computer Animation and Virtual Worlds 29*, 5 (2018). 2

[ARC22] ALVARADO E., ROHMER D., CANI M.-P.: Generating upper-body motion for real-time characters making their way through dynamic environments. In *Computer Graphics Forum, Proceedings of SCA* (2022), vol. 41, pp. 169–181. 6, 8

[Bad97] BADLER N.: Virtual humans for animation, ergonomics, and simulation. In *Proceedings IEEE Nonrigid and Articulated Motion Workshop* (1997), IEEE, pp. 28–36. 1

[BKB*12] BERNARDIN D., KADONE H., BENNEQUIN D., SUGAR T., ZAOUI M., BERTHOZ A.: Gaze anticipation during human locomotion. *Experimental brain research 223* (2012), 65–78. 6

[BKCZ09] BARTNECK C., KULIĆ D., CROFT E., ZOGHBI S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics 1* (2009), 71–81. 7

[CAC*22] CURTIS C., ADALGEIRSSON S. O., CIURDAR H. S., MCDERMOTT P., VELÁSQUEZ J., KNOX W. B., MARTINEZ A., GAZTELUMENDI D., GOUSSIES N. A., LIU T., NANDY P.: Toward believable acting for autonomous animated characters. In *Motion, Interaction and Games* (2022). 1

[CKB99] CHOPRA-KHULLAR S., BADLER N. I.: Where to look? automating attending behaviors of virtual human characters. In *Proceedings of the third annual conference on Autonomous Agents* (1999), pp. 16–23. 2

[DE09] DROLL J. A., ECKSTEIN M. P.: Gaze control and memory for objects while walking in a real world environment. *Visual Cognition 17*, 6-7 (2009), 1159–1184. 6

[DOA22] DELBOSC A., OCHS M., AYACHE S.: Automatic facial expressions, gaze direction and head movements generation of a virtual agent. In *International Conference on Multimodal Interaction* (2022), p. 79–88. 1

[EHSN19] EOM H., HAN D., SHIN J. S., NOH J.: Model predictive control with a visuomotor system for physics-based character animation. *ACM Transactions on Graphics (TOG) 39*, 1 (2019), 1–11. 2

[GBO*23] GOUDÉ I., BRUCKERT A., OLIVIER A.-H., PETTRÉ J., COZOT R., BOUATOUCH K., CHRISTIE M., HOYET L.: Real-time multi-map saliency-driven gaze behavior for non-conversational characters.

IEEE Transactions on Visualization and Computer Graphics (2023), 1–13. 3, 4, 7, 8, 10

[GD02] GILLIES M. F. P., DODGSON N. A.: Eye movements and attention for behavioural animation. *The Journal of Visualization and Computer Animation 13*, 5 (2002), 287–300. 2

[GPMJ13] GOMES P., PAIVA A., MARTINHO C., JHALA A.: Metrics for character believability in interactive narrative. In *International Conference on Interactive Digital Storytelling* (2013), pp. 223–228. 7

[HB05] HAYHOE M., BALLARD D.: Eye movements in natural behavior. *Trends in cognitive sciences 9*, 4 (2005), 188–194. 3

[HPV02] HOLLANDS M. A., PATLA A. E., VICKERS J. N.: "look where you're going!": gaze behaviour associated with maintaining and changing the direction of locomotion. *Experimental brain research 143* (2002), 221–230. 5

[IDP06] ITTI L., DHAVALE N., PIGHIN F.: Photorealistic attention-based gaze animation. In *International Conference on Multimedia and Expo* (2006), pp. 521–524. 2, 4

[KB23] KÜMMERER M., BETHGE M.: Predicting visual fixations. *Annual Review of Vision Science 9* (2023), 269–291. 2

[KBJ*] KÜMMERER M., BYLINSKII Z., JUDD T., BORJI A., ITTI L., DURAND F., OLIVA A., TORRALBA A., BETHGE M.: Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/. 3

[KSDG20] KRONER A., SENDEN M., DRIESSENS K., GOEBEL R.: Contextual encoder–decoder network for visual saliency prediction. *Neural Networks 129* (2020), 261–270. 2, 3, 4

[LHOK18] LOTH S., HORSTMANN G., OSTERBRINK C., KOPP S.: Accuracy of perceiving precisely gazing virtual agents. In *International conference on intelligent virtual agents* (2018), pp. 263–268. 7

[LM21] LOCKHOFEN D. E. L., MULERT C.: Neurochemistry of visual attention. *Frontiers in neuroscience 15* (2021), 643597. 2

[MdAM11] MIYASIKE-DASILVA V., ALLARD F., MCILROY W. E.: Where do we look when we walk on stairs? gaze behaviour on stairs, transitions, and handrails. *Experimental brain research 209* (2011), 73–83. 6

[MMA*23] MELGARÉ J., MAINARDI G., ALVARADO E., ROHMER D., CANI M.-P., MUSSE S.: How much do we pay attention? a comparative study of user gaze and synthetic vision during navigation. *Motion, Interaction and Games (Poster)* (2023). 4

[Mor70] MORI M.: Bukimi no tani [the uncanny valley]. *Energy 7* (1970), 33–35. 7

[MSB*22] MARTIN D., SERRANO A., BERGMAN A. W., WETZSTEIN G., MASIA B.: Scangan360: A generative model of realistic scanpaths for 360 images. *IEEE Transactions on Visualization and Computer Graphics 28*, 5 (2022), 2003–2013. 3

[NCRP16] NEOG D. R., CARDOSO J. L., RANJAN A., PAI D. K.: Interactive gaze driven animation of the eye region. In *Proceedings of the 21st International Conference on Web3D Technology* (2016), pp. 51–59. 2

[PCR*11] PETERS C., CASTELLANO G., REHM M., ANDRÉ E., RAOUZAIOU A., RAPANTZIKOS K., KARPOUZIS K., VOLPE G., CAMURRI A., VASALOU A.: Fundamentals of agent perception and attention modelling. *Emotion-Oriented Systems: The Humaine Handbook* (2011), 293–319. 2

[PDGea01] PURVES D A., GJ F. D., ET AL.: *Neuroscience*. Sinauer Associates, 2001, ch. Types of Eye Movements and Their Functions. 5

[PMG13] PEJSA T., MUTLU B., GLEICHER M.: Stylized and performative gaze for character animation. In *Computer Graphics Forum, Proceedings Eurographics* (2013), vol. 32, pp. 143–152. 4, 6

[Rat78] RATCLIFF R.: A theory of memory retrieval. *Psychological review 85*, 2 (1978), 59. 5, 6

[RM08] RATCLIFF R., MCKOON G.: The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation 20*, 4 (2008), 873–922. 2, 5

[RPA*15] RUHLAND K., PETERS C. E., ANDRIST S., BADLER J. B., BADLER N. I., GLEICHER M., MUTLU B., MCDONNELL R.: A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. *Computer Graphics Forum 34*, 6 (2015), 299–326. 1, 2

[RRHO23] ROTH N., ROLFS M., HELLWICH O., OBERMAYER K.: Objects guide human gaze behavior in dynamic real-world scenes. *PLOS Computational Biology 19*, 10 (2023), e1011512. 6

[RTT90] RENAULT O., THALMANN N. M., THALMANN D.: A vision-based approach to behavioural animation. *The journal of Visualization and computer animation 1*, 1 (1990), 18–21. 2

[SBR07] SPRAGUE N., BALLARD D., ROBINSON A.: Modeling embodied visual behaviors. *ACM Transactions on Applied Perception (TAP) 4*, 2 (2007), 11–es. 3

[SRJ11] STRASBURGER H., RENTSCHLER I., JÜTTNER M.: Peripheral vision and pattern recognition: A review. *Journal of vision 11*, 5 (2011), 13–13. 3

[SWHH90] SPECTOR R., WALKER H., HALL W., HURST J.: *Clinical Methods: The History, Physical, and Laboratory Examinations*. Boston: Butterworths, 1990. 4

[TGCL20] THOMAS N. D., GARDINER J. D., CROMPTON R. H., LAWSON R.: Look out: an exploratory study assessing how gaze (eye angle and head angle) and gait speed are influenced by surface complexity. *PeerJ 8* (2020), e8838. 2, 5, 6

[WSX*19] WANG W., SHEN J., XIE J., CHENG M.-M., LING H., BORJI A.: Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence 43*, 1 (2019), 220–237. 3

[YLNP12] YEO S. H., LESMANA M., NEOG D. R., PAI D. K.: Eyecatch: Simulating visuomotor coordination for object interception. *ACM Transactions on Graphics (TOG) 31*, 4 (2012), 1–10. 2

[ZZWT24] ZHANG Y., ZHANG T., WU C., TAO R.: Multi-Scale Spatiotemporal Feature Fusion Network for Video Saliency Prediction. *IEEE Transactions on Multimedia 26* (2024). 2, 3