Real-time Multi-map Saliency-driven Gaze Behavior for Non-conversational Characters

Ific Goudé, Alexandre Bruckert, Anne-Hélène Olivier, Julien Pettré, Rémi Cozot, Kadi Bouatouch, Marc Christie, Ludovic Hoyet

Real-time Multi-map Saliency-driven Gaze Behavior for Non-conversational Characters

Ific Goudé*, Alexandre Bruckert*, Anne-Hélène Olivier, Julien Pettré, Rémi Cozot, Kadi Bouatouch, Marc Christie, and Ludovic Hoyet

Abstract—Gaze behavior of virtual characters in video games and virtual reality experiences is a key factor of realism and immersion. Indeed, gaze plays many roles when interacting with the environment; not only does it indicate what characters are looking at, but it also plays an important role in verbal and non-verbal behaviors and in making virtual characters alive. Automated computing of gaze behaviors is however a challenging problem, and to date none of the existing methods are capable of producing close-to-real results in an interactive context. We therefore propose a novel method that leverages recent advances in several distinct areas related to visual saliency, attention mechanisms, saccadic behavior modelling, and head-gaze animation techniques. Our approach articulates these advances to converge on a multi-map saliency-driven model which offers real-time realistic gaze behaviors for non-conversational characters, together with additional user-control over customizable features to compose a wide variety of results. We first evaluate the benefits of our approach through an objective evaluation that confronts our gaze simulation with ground truth data using an eye-tracking dataset specifically acquired for this purpose. We then rely on subjective evaluation to measure the level of realism of gaze animations generated by our method, in comparison with gaze animations captured from real actors. Our results show that our method generates gaze behaviors that cannot be distinguished from captured gaze animations. Overall, we believe that these results will open the way for more natural and intuitive design of realistic and coherent gaze animations for real-time applications.

Index Terms—Gaze behavior, simulation, animation, neural networks, eye-tracking data, dataset.

1 Introduction

Realistically animating virtual characters in real-time applications remains a challenging issue, due to the sheer number of elements that contribute to their final realism. In this paper, we focus on gaze animation which, as identified by several authors (*e.g.*, [1]), plays an important role in making virtual characters come alive. E.g., gaze provides information about what characters are looking at, takes part in verbal and non-verbal communication, conveys information about the emotional state of the characters. More importantly, with the increase of VR applications, we expect that providing virtual characters with realistic eye-gaze animations will have an important positive impact on user experience and immersion.

In this context, addressing the challenges to create realistic gaze animations means: i) at the lowest level, generating eye-head motion trajectories that match human ones, ii) at a higher level, generating gaze sequences that exhibit realistic kinematics and oculomotor properties and iii) driving the attention of characters coherently with the visually important elements of the scene. To our knowledge, no real-time method is able to address all of these challenges automatically as of today, without users providing semantic information about the scene.

- ** both authors have contributed equally to the submission*
- *Ific Goudé, Anne-Hélène Olivier, Julien Pettré, Kadi Bouatouch, Marc Christie and Ludovic Hoyet are with Inria, Univ Rennes, CNRS, IRISA.*
- *Alexandre Bruckert is with Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France. E-mail: alexandre.bruckert@univ-nantes.fr*
- *Rémi Cozot is with Littoral Opal Coast University*

Previous methods tackle these problems in two different ways that can be distinguished into object-based or image-based approaches. In object-based approaches, the character's gaze is animated in the direction of interesting objects in the scene, where level of interest is defined by hand-crafted features or semantics (*e.g.*, [2], [3]). In image-based approaches, the gaze is animated in the direction of attractive areas in the character's field of view, mainly based on simple heuristics to simulate the human visual attention (*e.g.*, [4]). Image-based methods are more generic as they do not require to manually define the semantic of the scene, but still have difficulties in predicting the human visual attention, which is influenced by multiple factors (*e.g.* task, memory, personal interest). Moreover, they often suffer from high computational requirements, which prevents their usage in interactive applications.

We therefore propose a novel image-based method that leverages recent advances in several distinct areas related to visual saliency, attention mechanisms, saccadic behavior modelling, and head-gaze animation techniques. Our approach (detailed in Section 3) articulates these advances to converge on a multi-map saliency-driven model which offers real-time realistic gaze behaviors for non-conversational characters, together with additional user-control over customizable features to compose a wide variety of results. More specifically, our method is built on three main components that provide high visual realism and coherency with the scene: a *saliency model*, a *saccadic model* and an *eye-head animation model*. The *salicency model* is in charge of assessing which elements in the character's field of view are more likely to attract its visual attention, solely based on the content of the scene. Then, considerations about the human

eye kinematics are handled by the *saccadic model*, which role is to reproduce natural oculomotor characteristics (*e.g.*, duration of eye fixations, orientations and amplitudes of eye saccades, etc.). It also enables users to easily customize novel features to synthesize specific gaze behaviors in a common framework without directly specifying how to control the character's gaze. Finally, the *eye-head animation model* controls the eyes and head motion, while ensuring the realism of the kinematic variables, which plays a key role in the realism of the animation. We illustrate the results of our method through several examples, presented in the supplementary video, where virtual characters freely explore their environment. We also demonstrate the ability of our model to include customized behaviors through two examples: a) progressively including an horizontal bias (*i.e.*, the tendency for the gaze to rest on the horizon line when there is no specific task to perform), and b) including an attraction map to influence the gaze toward a specific area while retaining some of the stochasticity provided by the other parts of the model.

We also present two evaluations of our method. First, an objective evaluation (Section 4) measures the similarity of our gaze simulations to real human gaze when exploring virtual environments, using different metrics. For this purpose, we conducted a user study to collect ground truth eye-gaze activity over 50 participants. We then conducted an online subjective evaluation of our model (Section 5) to assess the perception of the realism of the synthesized gaze of a virtual character. Our results show that our model is able to generate eye-gaze animations perceived to be visually similar to gaze animations captured from real actors.

To summarize, our main contributions are the following:

- We propose a novel multi-map saliency-driven framework for simulating realistic, interactive, and easily customizable gaze behaviors for virtual characters. This framework combines for the first time in a unique real-time implementation recent advances regarding visual saliency, saccadic models, and eye-head coordination. Moreover, a C++ implementation of our framework for Unreal Engine 5 is made available online for the community[1].
- We propose a novel manner of enabling users to easily customize dedicated features to synthesize gaze behaviors in this common framework, without requiring complex hand-crafted or scripted constraints to control the character's gaze.
- We propose both objective and subjective evaluations of this framework. We find that this twofold evaluation is particularly relevant in the context of gaze simulation in virtual reality environments.
- We captured and make available a novel eye-tracking dataset including the eye-gaze activity and head movements of 50 participants freely exploring three different virtual scenes in a variety of conditions. The dataset is also made available for the community, along with our framework implementation[1].

Overall, we believe that our results will open the way for natural and intuitive manners of designing and generating realistic gaze animations for real-time applications.

1. https://github.com/igoude/saliency-driven-gaze

2 Related work

In this section, we present relevant methods for simulating gaze behaviors of virtual characters, either relying on visual attention or saccadic models, as well as techniques for animating the eyes and head of such characters. We also refer the reader to the review of Ruhland et al. [1] *w.r.t* eye-gaze for a more comprehensive and detailed review.

2.1 Saliency models

When animating virtual characters, the environment plays a crucial contextual role in terms of gaze behavior. The first model of visual attention for a virtual character was proposed by Khullar and Badler [5], who argue that such a model is useful to determine the ergonomics of virtual environments, hence improving the behavior of avatars in cyber-chats or virtual characters in video games. They propose a task-based model that reacts to the exterior environment, guiding the locomotion and controlling the eye-head motion of the character. The visual attention is directed thanks to a list of endogenous (*e.g.*, user-defined tasks, scene graph queries) or exogenous (*e.g.*, peripheral motion, saliency) factors, where the visual saliency is given by the contrast of the image in the character's field of view.

Later, Peters et al. [6] present a review of existing visual attention models for virtual characters. They draw up a comparison table that refers to seven methods [4], [5], [7], [8], [9], [10], [11] compared in terms of key factors (*e.g.*, top-down, bottom-up, movable character). The authors discuss the evaluation of such methods, that may rely on quantitative comparisons (against human gaze data) or subjective studies (by judging the realism of gaze behavior). They distinguish two families of techniques in virtual-world applications, object-based and image-based methods (with several hybrid approaches that combine both).

Object-based methods rely on scene queries from an omnipotent point of view. Gillies and Dodgson [2] improve the model of Khullar and Badler [5] by driving the character's gaze depending on its destination and its head orientation in addition to performing requests on objects in the scene. The gaze behavior model works as a state machine that switches between scene requests, objects of interest in the character's field of view, the forward direction and a random direction. In contrast, the method of Oyekoya et al. [3] relies on the distance, the velocity and the eccentricity of objects relative to the character. They compare the visual realism of their method and show that it results in more realistic gaze animations than models of static eyes and random gaze, but is less realistic than a live tracking of an actor's eyes. Recently, Ağıl and Güdükbay [12] proposed a method for pedestrian crowds including group effects and personality, succeeding in simulating gaze behaviors for groups of people comparable to real street sequences.

Image-based methods only depend on the rendered image of the virtual environment from the character's field of view. Courty et al. [8] propose a saliency model dedicated to obstacle avoidance for autonomous characters, where the saliency is given by a combination of a Gabor filtering (*i.e.*, high frequencies) of the colored image, the depth buffer and a centered 2D Gaussian map. Pixels with high values in the processed image are defined as salient and their position

is projected back onto the 3D world. The eyes and head of the character are then directed to focus on the salient point in the scene. Meanwhile, Peters and O'Sullivan [7], [13] used both scene rendering (to compute the saliency) and scene queries (to obtain object velocities) to propose a more faithful visual attention model. Their saliency map (16×16 pixels) is based on psychophysical models and is computed at each frame. The maximum salient area is detected and an Inhibition of Return is added on the image or to objects to avoid always looking at the same area. A feedback loop is then added to simulate the characters' memory.

Finally, Itti et al. [4] propose a complex model of visual attention for autonomous characters. The model focuses on bottom-up rather than top-down visual attention and is adapted to temporal changes (*e.g.*, moving objects, light blinking). Their saliency map, that relies on their computational modeling of visual attention [14], is combined with a task-relevance map to guide the gaze of the character. Regarding the animation, the eye and head movements follow psychophysical rules that are simple and efficient. However, this method cannot be performed in real-time and the entire animation requires to be preprocessed.

2.2 Saccadic models

To produce more realistic movements, several works explored the question of simulating gaze using saccadic models. Lee et al. [15] first propose a statistical model derived from eye-tracking data for animating a character's eyes. Their subjective study suggested that a conversational agent animated thanks to their statistically-derived model is perceived as more interested, engaged, friendly and lively than an agent with either random eye animations or without any eye animation. Pelachaud and Bilvi [16] then enforced more variety by generating the audio and the face animation of the character, including eye gaze, given a text it has to say. Five gaze parameters are empirically defined, which control the duration of different gaze states (*e.g.*, mutual gaze state between two characters, speaker state, listener state) and then modify the behavior of the characters.

Following the idea of gaze directed by mutual interactions between two characters, Satogata et al. [17] propose a model directed by an energy function motivated by three parameters: the desire to look, the mutual gaze hesitation, and the mutual gaze stress. Each parameter can be weighted to adjust the character's personality. They showed that a shy robot animated with their model effectively starts averting its gaze and turns away when being looked at by a human.

2.3 Animation models

To increase visual realism, another family of approaches focus on jointly modeling eyes and head trajectories during gaze shifts. For instance, given the current positions and orientations of the character's eyes and head as well as a gaze target to reach, Andrist et al. [18] compute deterministic trajectories of eyes and head that rely on variables from neurophysiological studies (*e.g.*. head latency, eye and head velocities, oculomotor range). Steptoe et al. [19] also explore specific animation details such as eyelid animation by separating the lid saccades (happening during a gaze shift) from blinks (initiated by voluntary or involuntary stimuli) and present a parametric model derived from literature in ophthalmology and psychology.

Definitions	
Gaze	Movement of eyeballs, eyelids, head and more generally the entire body to look at something or someone.
Saccade	Rapid shifts in eye position that centre the gaze on targets of interest [1].
Fixation	Short period of time between two saccades where the gaze focuses on a specific area.
Eyes and head movement	Coordination of the body parts to distribute angles and courses in the gaze behavior.
Blink	Spontaneous, voluntary or reflexive rapid eyelid movement.
Visual attention	Scene exploration directed by both image-based saliency cues (bottom-up) and task-dependent cues (top-down) [14].
Saliency	Stimulus conspicuity, visually attractiveness.
Saliency map	Explicit two dimensional topographical map that encodes saliency at every location in the visual scene [14].

TABLE 1: Definitions of terms used in the description of the method.

The same year, Peters and Qureshi [20] propose to animate gaze shifts using two components: a eye-head controller and a blinking controller. The virtual character's eyes and head are animated toward a given target location in the virtual environment while the animation of eyelids simulates gaze-evoked blinking (*i.e.*, a specific category of behaviors related to gaze shifts). A number of experiments were conducted to address the perception of blinking strategies, naturalness of eye-head movement ratios, and differences between horizontal and vertical gaze shifts.

Thanks to the grow in deep learning techniques, Klein et al. [21] propose a data-driven gaze animation method using Recurrent Neural Networks (RNN). The network is trained to learn the different eye-gaze constraints imposed by the posture (*e.g.*. standing, sitting, lying), and the resulting animations are perceived to be more natural than those using procedural gaze animations. Unfortunately, for now the network is effective only for gaze pursuits (*i.e.*, following moving targets) and does not handle gaze shifts.

2.4 Summary

To summarize, models for simulating gaze behaviors have drastically improved over the last two decades. Nevertheless, no method has achieved to date a realistic simulation that can simultaneously be performed in real-time, react to scene events or user interactions, and be easily customized to generate a variety of gaze behaviors. We therefore propose the first real-time gaze simulation method that provides a realistic and contextual gaze behavior by conjugating data-driven models of visual attention, saccadic movement, and eye-head animation. To the best of our knowledge, our method is the first to offer a fully-automatic simulation of gaze behaviors by integrating these different models into a unified framework, and achieving this level of realism while still providing control and customization over specific visual features, designed to incorporate visual biases and semantic information.

3 Saliency-driven gaze animation model

In this section, we present our saliency-driven gaze animation model. We first present the overall workflow, followed

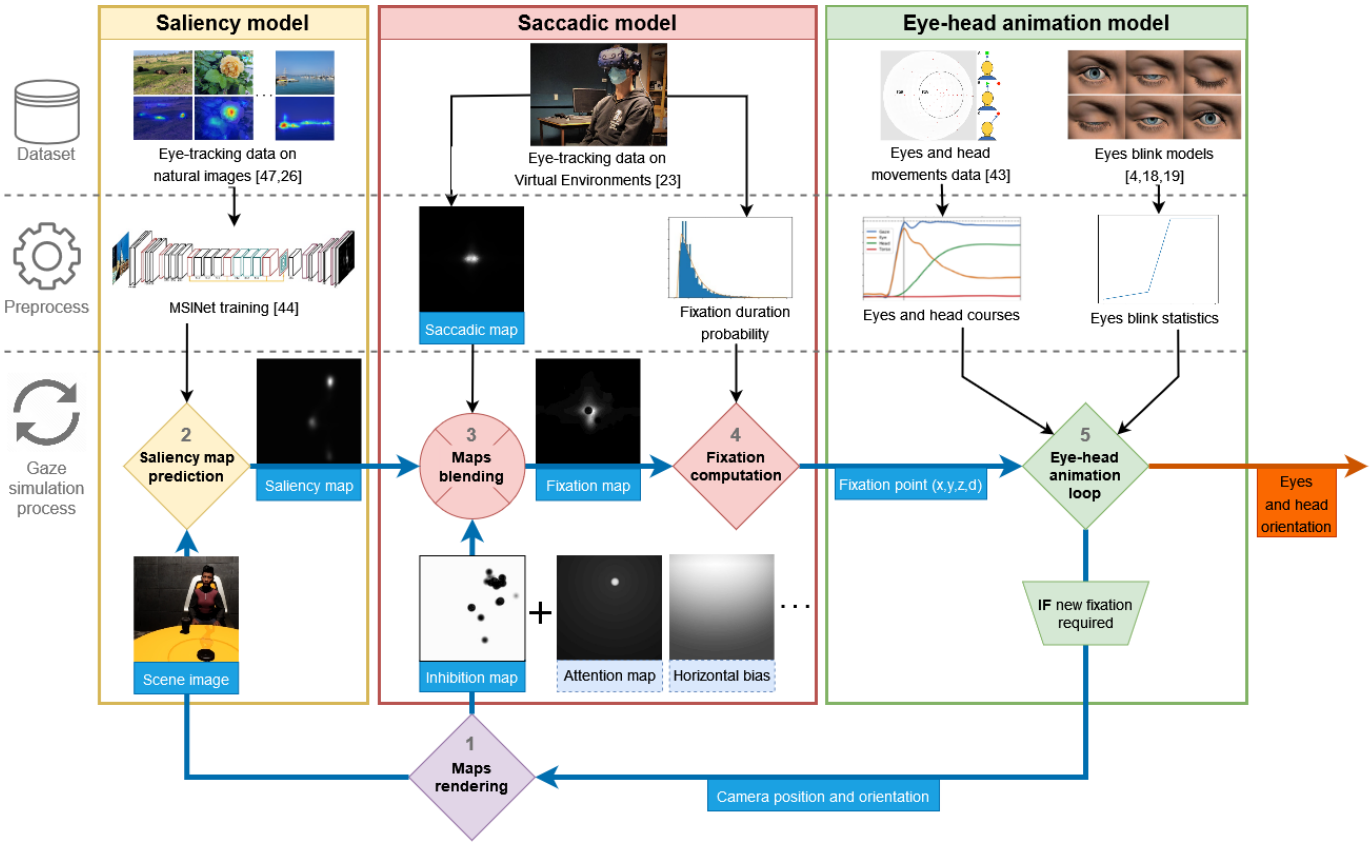


Fig. 1: The flowchart of our method describing the five-step process (numbered blocks): 1) Rendering the image in the character's field of view. 2) The image of the scene is passed through a visual saliency model, which outputs an eye-fixation probability for each pixel. 3) The predicted saliency map is combined with several human oculomotor biases, and merged into a fixation distribution. 4) The position and the duration of the fixation are randomly sampled, using the computed spatial fixation distribution and predetermined fixation duration distribution. 5) Given the character's current eye and head orientations, its gaze is animated toward the new fixation point. When the duration of the fixation is reached, the process is reiterated from step 1.

by implementation details and results obtained using our model. Additional details for reproducing our framework are given in the supplementary material. Important terms and definitions used in our method are presented in Table 1.

3.1 Workflow

The *Gaze Simulation process* that controls the activity of a character's gaze is represented transversely on the lower part of Figure 1. This process takes as input the image of the scene as perceived by the virtual character (*i.e.* the scene rendered from the character's point of view) and outputs its gaze animation to look at a given point in the scene (*i.e.* the new fixation point). The computation of this fixation point given the character's point of view of the scene results from the five following steps (numbered blocks in Figure 1):

1) When a new fixation is required, an **image of the scene is rendered** from the character's point of view, and a new iteration of the gaze simulation is performed.
2) The **saliency map** corresponding to the image of the scene is predicted using a deep learning approach. This determines the gaze attraction level for each point in the image, *i.e.* the entities which are likely to be gazed at in the character's field of view.
3) The **saliency map** is then combined using a weighted product with the **saccadic map** (*i.e.*, the map of plausible fixation point transitions), and the **inhibition map** (*i.e.*, preventing the gaze to go back to the same fixation points). This results in the **fixation map**, which corresponds to the probability of producing a saccade in a particular direction. At this step, **customization maps** can also be combined to influence the resulting **fixation map**, providing means to bias the fixation map to reproduce various phenomenons *e.g.*, attracting the gaze toward objects with specific semantics, as presented in Section 3.3. The combination operation is a weighted average of the different maps, which enables to generate various gaze behaviors.
4) The **position and duration of the new fixation** are randomly sampled, using the **fixation map** and the predetermined fixation duration distribution.
5) An **animation of the character's eyes and head** is computed to reach this new fixation point, simultaneously generating eye blinks in a consistent way.

Our *Gaze Simulation process* is built on three main com-

ponents (presented vertically in Figure 1) that are strongly data-driven: a *saliency model*, a *saccadic model* and an *eye-head animation model*:

1) The *saliency model* takes as input the image of the scene rendered from the point of view of the character and generates a saliency level (a scalar value) for each pixel of this image. This **saliency map** is predicted by a network called MSI-Net which is itself trained on a dataset combining images (of natural scenes) and eye-tracking data (filtered fixations of viewers). While there might be some domain-shift issue due to the current model being trained on 2D natural images and applied to 3D rendered scenes, this approach has already been used with success on rendered 3D objects [22]. Additionally, the scenes that we use are quite sparse in objects and details. This means that visual attention will be mostly drawn by low-level features, such as color or luminance contrasts, which are very well extracted by deep saliency models. These elements are further discussed in Section 6.
2) The *saccadic model* captures the probabilities of saccade kinematic variables (saccade amplitude and orientation) as well as duration, which enables us to generate realistic transitions between different eye fixation points. This model relies on a **saccadic map** and a probability of fixation duration, which can be extracted from existing eye-tracking datasets.
3) The *eye-head animation model* controls the eyes and head motion between the fixation points in a realistic way. Given the position of the character and a fixation point, it guarantees that the angles formed by the body, head and eyes are realistically distributed and that the kinematic variables follow realistic courses. It also controls the eyelid animation so as to generate realistic eye blinks.

3.2 Implementation details

This workflow was implemented in Unreal Engine 5, using MetaHuman characters to apply our gaze simulation model. The *saliency model* is trained on natural images and implemented in C++ with the Libtorch library. The saccadic map and the fixation duration probability used by the *saccadic model* are extracted from the data gathered by David *et. al.* [23].

The results to be presented in this section were computed using a Dell Precision 7528 (CPU intel CORE i7 vPro 8th Gen, GPU NVidia Quadro M2200, 32GB RAM). Our model has a low impact on performance as the gaze simulation process occurs only once at the end of each fixation, *i.e.*, every 100 to 400 ms. The computation time regarding the scene rendering depends on both the complexity of the scene and the rendering quality expected. However, a 256×256 image is required by the *saliency model*, which is rendered very quickly with most consumer-grade computers. The prediction of the **saliency map** has a low processing time and takes on average 6 ms to compute. Furthermore, this process is threaded to prevent freezing the main game thread. The computation of the **inhibition map**, the **customization maps**, and the *maps blending* process all rely on simple pass shaders. We also applied the following coefficients in the weighted product to the **saliency**, **saccadic** and **inhibition maps**: 0.9, 0.85, and 0.75. These weights were selected empirically to put enough emphasis on each map while leaving enough different attractive areas so that the character can explore the scene instead of getting stuck in a local maximum. The fixation computation includes a trivial loop over 20 pixels, and a unique ray-casting to determine the fixation point. Finally, the eye-head animation loop is computed in the main animation thread.

All the technical details *w.r.t* the reproduction of the different components of our model are provided in the supplementary material.

3.3 Results

To illustrate our method, we designed four populated virtual scenes corresponding to different everyday-life situations: a lobby room with two characters sitting at a table, a waiting room, a bar with a football match broadcast on TV, and a poker table (see Figure 2). Complexity ranged from 46k triangles for the simplest scene (the poker table) to 453k triangles for the most complex one (the waiting room). Despite differences in scene complexity, none of the scenes suffered from rendering issues, such as lag or visual artefacts. Note that the scenes do not contain any sound as our method only considers the visual aspect, although this could be an interesting direction for future work as discussed in Section 6.

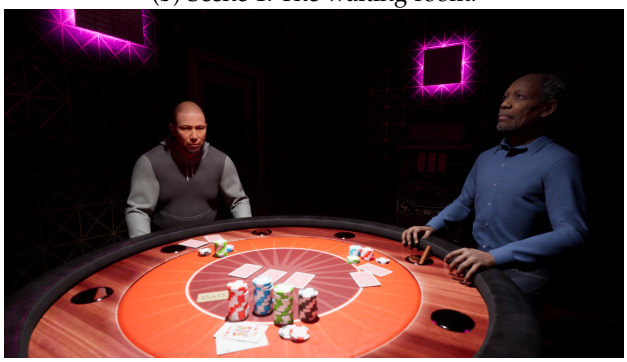
Without adding any **customization map**, our model generates eye-gaze animations of virtual characters that are freely exploring their environment, as displayed first in our supplementary video (Free Exploration Results). Such characters are very common in video games and other interactive applications (*e.g.* pedestrians, customers), and often suffer from a lack of eye-gaze activity. We then present examples using additional **customization maps**, which enable users to easily control and synthesize specific gaze behaviors in a common framework. In particular, we present hereafter two specific examples (which are also presented in the supplementary video), and discuss their benefits and generalization over other solutions.

Horizontal bias (T_{HB}): Without specifying additional behaviors, characters are left free to explore the whole scene, as one would when he/she discovers a new place. However, it is also known than when no particular task has to be performed, the gaze tends to progressively rest on the horizon line, rather oriented in the body forward direction. This leads over time to an oculomotor bias known as *horizon bias* [24], which seems to be mainly explained by the fact that in the resting position the head is straight and the eyes look at the horizon. As such a bias is not captured by the **saccadic map**, we propose to model it using a custom **horizontal bias map** T_{HB}, which increases the probability of producing a fixation on the horizon line in front of the character. We use this map by increasing its weight over time, letting the character fully explore the scene first, and progressively increasing the probability of going into a resting posture (*i.e.* looking at the horizon). Figure 3a shows the contribution of this map and how it directs the gaze.

(a) The lobby.

(b) Scene 1: The waiting room.

(c) Scene 2: The bar.



(d) Scene 3: The poker table.

Fig. 2: Populated scenes used to present our results (Section 3.3) and for the user experiment (Section 4.1).

(a) The **horizontal bias map** T_{HB} increases the probability to look at the horizon instead of the ground.

(b) The TV in the back of the bar is very attractive semantically, which is enforced using the **attention map** T_{att}.

Fig. 3: Our **customization maps** represented as Heat-maps.

Attention map (T_{att}): Similarly, in some cases we want a character to look at a specific object, *e.g.*, that might have a low saliency level (*i.e.* not a particular visual appearance) but a high semantic importance in the current situation or at a current time. As an example, the football match that is broadcasted on the TV in the *bar scene* is very attractive by its semantic, but not by its saliency. Such behaviors can be brought about by using a custom **attention map** T_{att}, which defines a mask around specified objects of interest. In particular, it provides the benefits of eliciting the gaze to be directed toward this area, while retaining some of the stochasticity provided by the other parts of the *saccadic model*. The effect of this map is presented in Figure 3b.

Discussion: The use of these **customization maps**, and the level of control users have over them, makes our approach generic to many use cases. In particular, it enables users to design specific behaviors in a common framework, without relying on hard-coded or scripted constraints directly controlling the gaze direction. For instance, while it would be possible to manually force the character to look at the TV in the *bar scene*, it would simultaneously be difficult to manually specify for how long the TV should be looked at, and when to switch temporarily between different elements of the scene. Our model enables such behaviors to naturally emerge, as it retains the stochasticity provided by the other parts of the model, *e.g.*, enabling the character to automatically observe salient other characters for a few seconds before looking back at the TV. Moreover, scenarizing specific events and character behaviors becomes extremely easy, as it only requires to temporally control the respective weight of the different **customization maps**. Finally, while we only presented in this section two examples that we use for synthesizing our results, we believe that several other types of **customization maps** could be included in the future to augment the repertoire of potential behaviors, *e.g.*, based on specific animation information such as facial expressions, on optical flow information, on a sound location to attract the character's attention, etc.

The following sections will then focus on evaluating the resulting gaze animations, both objectively and subjectively.

4 Objective Evaluation

In this section, we present an objective evaluation of our method to evaluate the similarity between our gaze behavior model and human gaze behavior. For this purpose, we

conducted a user study to collect ground truth eye-gaze activity, which we then compare to simulations generated by our model in the same visual conditions.

4.1 Eye-Gaze Data Collection

This section briefly presents the experimental details for the collection of the ground truth eye-gaze activity, and we refer the reader to the supplementary material for the full experimental description.

Fifty participants, who volunteered for our experiment, were equipped with an HTC Vive Pro Eye Head Mounted Display, which includes a built-in eye-tracker used as an assessment tool of saccadic eye movement [25]. Participants were immersed in the four virtual scenes presented in Section 3.3 (see Figure 2). The *lobby* room was used for training purposes, as well as for checking eye-tracking accuracy between trials, while the waiting room, bar, and poker table scenes were used for collecting ground truth data. For each of these three scenes, we also included combinations of the two following situations:

1) *Populated*: the scene either included virtual characters or not. Our objective was to account for differences in gaze behaviors in the presence of other characters, as faces are known to have a high level of attraction.
2) *Event*: the scene either included a particular event or not. The event was specific to each scene: the door of the waiting room suddenly opening, a red exit light suddenly flashing in the bar, chips falling from the slot machine in the poker room. Our objective was to account for sudden gaze behaviors when an event occurs, and to demonstrate the adaptability of our model in such situations.

In total, participants performed 12 trials in random order (3 *Scenes* $\times$ 2 *Populated* $\times$ 2 *Event*), which they freely explored for 30 seconds each while seated (they were not allowed to get up from the chair). During the experiment, we used the SRanipal SDK to collect participants' eye-tracking data at 90Hz: the participant's head (HMD) position and rotation in the virtual space, the gaze direction vector (*i.e.*, the combined direction of head and eye in the virtual space) and the eye openness (from 0: closed to 1: opened). This data was processed according to the procedures described in the supplementary material to extract the scene saliency maps used in the following evaluation.

4.2 Simulation Creation

To efficiently compare gaze simulations produced by our method with real gaze behaviors collected during our experiment, we created virtual situations matching those observed in the experiment. For each virtual scene, a virtual character is positioned at the same location than participants, and oriented toward the same direction, with its head at the average position of the participants' head. It was also animated using an idle animation to resemble the body gestures of participants when exploring the scenes. To generate more variety in the simulations, we randomly initialized the direction of the character's gaze in a cone of 90° aligned with the character's head forward vector.

In this evaluation, we are interested in the similarity between real gaze behaviors and those produced by our model. We therefore only consider the generated saccades and fixations (*i.e.* saccade amplitude, angle, and fixation location), which result from the *saliency and saccadic models*, and not the *eye-head animation model*, which only influences the final animation of the character. We therefore performed an ablation study to evaluate the relative influence of these two components using four different configurations:

1) *Random*: the **saliency**, **saccadic** and **inhibition maps** are not considered, resulting in a random gaze simulation.
2) *Saccadic*: only the **saccadic map** is used, saliency and inhibition maps are deactivated.
3) *Saliency*: only the **saliency map** is used, saccadic and inhibition maps are deactivated.
4) *Complete*: we used the complete model as presented in Section 3, with the same weights than for our results ($w_{saliency} = 0.9$, $w_{saccadic} = 0.85$, $w_{inhibition} = 0.75$).

A slight **horizontal bias** ($w = 0.01$) is however applied in each configuration to prevent the character's gaze from wandering too far to the sides or behind.

For each of the 12 conditions (the same as presented to participants and described in Section 4.1), in all of the four configurations of our model described above, we performed a 30-second simulation (same duration as our ground truth eye-gaze activity collection). For each fixation generated by the simulation, we collected its duration and location in the scene, the character's head position and rotation, the character's gaze direction, and the character's eye openness. The fixation distributions extracted from the eye-tracking experiment were then compared with those generated by our simulations as explained below.

4.3 Objective metrics

For our objective evaluation, we use two families of metrics. One is derived from the evaluation of visual saliency models by measuring the similarity between the fixation distributions of the ground-truth and of our simulations. The second compares the ground-truth and the simulated scanpaths (*i.e.* the sequences of fixations). These objective metrics rely on 3-dimensional saliency maps of the scenes. They were adapted from metrics used in the 2D image domain, and are detailed in the supplementary material.

Saliency-based metrics: In this first family of metrics, we evaluate the similarity between the fixation distributions in two ways: 1) comparing the ground-truth saliency map S to the simulated saliency map $\hat{S}$ (as in CC and KLD metrics described below), and 2) comparing the raw ground-truth binary fixation map F (*i.e.* a map where each pixel value is either one if it was fixated, and zero otherwise) to the simulated saliency map $\hat{S}$ (as in NSS and AUC metrics described below). In the following, we provide a short description of the different metrics used:

- The Pearson's correlation coefficient (CC) evaluates the linear relationship between S and $\hat{S}$. It outputs values between -1 (perfect negative correlation) and 1 (perfect positive correlation). It is symmetric, and

thus does not distinguish between false positives (*i.e.* a predicted salient area where no fixation occurs experimentally) and false negatives (*i.e.* a predicted non-salient area where fixations occur).

- Kullback-Leibler divergence (KLD) is used to measure how the probability distributions of S differs from $\hat{S}$:

$$KL(S, \hat{S}) = \sum_i S_i \log \left(\varepsilon + \frac{S_i}{\varepsilon + \hat{S}_i} \right) \quad (1)$$

where i iterates over the pixels of the maps and ε is a regularization constant. The value of ε will affect how pixels with a prediction of zero will be penalized (as KLD is very sensitive to zero-values, and thus highly penalizes false positives). Identical maps will score very close to zero, and the score increases as the maps differ. The upper-bound for the metric depends on the size of the maps and the chosen value of ε.

- Normalized scanpath saliency (NSS) averages the values of the normalized and centered simulated saliency map $\hat{S}$ where fixations F occur. The chance level is 0, negative values indicate anti-prediction, and the higher the value, the better the prediction. This metric is particularly sensitive to false positives.
- Area under curve (AUC) metrics rely on the interpretation of the saliency map as a binary classifier of which areas are fixated or not. Similarly to NSS, it evaluates the saliency at the locations of ground-truth fixations. Based on a set of thresholds, the Receiver Operating Characteristic (ROC) is inferred to compute the AUC. As different ways of computing true and false positive rates exist, we rely on the implementation of Judd et al. [26].

Furthermore, two evaluations are performed using the saliency-based metrics:

1) *Global*: all fixations are aggregated, regardless of when they might occur, and a unique global score is calculated.
2) *Dynamic*: the fixations are pooled in 100ms time windows and a score is calculated by time window. All of these scores are finally averaged over time.

Scanpath metric: For our second family of metrics, scanpaths are compared following the approach of David et al. [23] in their benchmark for visual attention models on omnidirectional videos. We used the MultiMatch metric proposed by Jarodzka et al. [27], which confronts the simulated fixations of our model to the recorded fixations of a human observer for the same stimulus, based on several characteristics: the spatial proximity between fixations, the orientation and amplitude differences between saccades, and the temporal proximity. More specifically, we used the implementation of the Salient360! benchmark [23] using the orthodromic distance on the viewing sphere instead of the euclidian distance. Recall that, whether for ground-truth data or simulated data, the position of the observer's head is almost fixed and the exploration of the scene is 360°. Moreover, the observed scenes are dynamic, so metrics used for the evaluation of 360° videos are easily adaptable to our experimental conditions. Since the spatial differences between the fixations are already measured by the saliency metrics, and since the duration of the fixations are sampled from real fixation duration distributions, we only consider the saccadic length and angular similarities given by the MultiMatch metric. These two measures are then normalized and averaged together to produce a similarity score ranging from 0 (perfect similarity) to 1 (high dissimilarity).

4.4 Results

Ablation study: We compare the four different configurations of our model, as described in Section 4.2 (*i.e. Random*, *Saccadic*, *Saliency*, *Complete*), to evaluate the contribution of each module to the overall results. Table 2 presents the objective scores averaged over all the scenes. To provide some information about inter-participants variability, we also include the infinite humans baseline as an upper bound [28], which is computed by evaluating the saliency and fixation maps inferred from half of the participants against the other half, cross-validated over subjects.

As expected, on the saliency metrics the *Saccadic* configuration results in the worst similarity since the content of the scene is not taken into account: the gaze stays stuck in a very limited area of the scene with little exploration. In contrast, the *Random* configuration enables a wider exploration of the environment, resulting in higher scores than *Saccadic* regarding saliency metrics. However, the *Random* configuration shows very poor KLD scores due to the high number of false positives, *i.e.* points in the scenes wrongly predicted as salient. Interestingly, the CC scores are however quite high, especially in the global case. It can be explained by both, the human tendency to explore scenes more horizontally [24], and the fact that saliency metrics are not yet well-defined in the context of omnidirectional stimuli [29], thus allowing this kind of simple baselines to perform on a similar level with specialized models.

On the scanpath similarity metric, the *Saccadic* and *Complete* configurations show the best scores as they both include human oculomotor biases. As expected, the *Random* method performs quite poorly as simulations resulting from this configuration exhibit large and sudden saccades. Although the *Saliency* configuration has no oculomotor constraint, it still performs relatively well on this metric. Indeed, the *Saliency* configuration tends to get stuck in local saliency maxima, simulating short saccades as humans would do.

Overall, the *Complete* configuration provides a good trade-off between the spatial precision of predicted fixations, as shown by the global and dynamic saliency scores, and the human-like characteristics of saccades, promoted by the scanpath similarity metric.

Attention maps on events: One of the shortcomings of our non-customized model is the way it responds to events. As explained in Section 4.1, we designed events to take place in half of the situations experienced by participants. These events are not necessarily very salient from an image point of view (*e.g.*, the chips falling from the slot machine in the poker scene), but are semantically very rich and induce a high degree of visual congruency between human observers. To address this issue, we demonstrate the value of including an additional **attention map** linked to the events

	Global Saliency				Dynamic saliency				Scanpath similarity
Method	CC ↑	NSS ↑	KLD ↓	AUC ↑	CC ↑	NSS ↑	KLD ↓	AUC ↑	Vector Similarity ↓
Infinite humans [28]	0.959	5.524	1.013	0.955	0.493	8.381	4.186	0.713	0.301
Random	0,505	2,612	5,001	0,859	0,214	2,059	12,521	0,616	0.497
Saccadic	0.183	1.186	7.950	0.820	0.065	0.794	20.041	0,453	**0.338**
Saliency	0,390	2,618	1,787	0,936	0,294	**4,457**	**4,978**	0,588	0.398
Complete	**0.592**	**3.652**	**1.099**	**0.937**	**0.357**	4.325	5.517	**0.694**	0.343

TABLE 2: Using several metrics (detailed in Section 4.3), we compare four different configurations of our model (*Random*, *Saccadic*, *Saliency*, *Complete*) to evaluate the contribution of each module to the overall results. The infinite humans baseline [28] is also provided as an upper bound reference, and is computed by evaluating the saliency and fixation maps inferred from half of the participants against the other half, cross-validated over subjects. The presented scores are averages over all the scenes of the experiment. ↑ means higher score is better, while ↓ means lower score is better.

	Scene 1 (Waiting room)				Scene 2 (Bar)				Scene 3 (Casino)			
Method	CC ↑	NSS↑	KLD ↓	AUC ↑	CC ↑	NSS ↑	KLD ↓	AUC ↑	CC ↑	NSS↑	KLD ↓	AUC ↑
Infinite humans [28]	0.718	12.056	2.875	0.913	0.780	11.744	2.636	0.925	0.757	12.825	2.279	0.911
Saccadic	0.059	0.583	21.030	0.420	0.062	0.591	20.935	0.441	0.127	1.262	19.201	0.567
Saliency	0.200	2.502	9.810	0.455	0.089	1.650	13.30	0.406	0.137	1.711	6.141	0.532
Complete	0.261	3.143	7.687	0.618	0.127	1.845	11.715	0.547	0.136	1.901	7.427	0.546
Complete + attention map	**0.316**	**6.515**	**6.929**	**0.864**	**0.348**	**5.849**	**7.086**	**0.711**	**0.651**	**8.482**	**2.879**	**0.798**

TABLE 3: We evaluate different settings of our model (*Saccadic*, *Saliency*, *Complete*, *Complete with an additional attention map*) on scenes containing scripted events, and more specifically on the frames where those events occur. Those settings are also confronted with the infinite human baseline [28] against several metrics detailed in Section 4.3. ↑ means higher score is better, while ↓ means lower score is better.

to attract the gaze of the character to this area, as described in Section 3.3. Table 3 shows that this *Complete + attention map* configuration leads to a significant improvement in all of the saliency scores over the frames when an event occurs. Scanpath metrics are not evaluated here as we focus on assessing fixation positions related to the event positions themselves, and not saccadic movements in this context. However, the relatively high KLD scores seem to indicate that the attention maps we used were probably not focused enough on the specific location of the event, thus leading to simulated fixations to occur outside this very salient area.

5 Subjective evaluation

In addition to the objective evaluation, we performed a subjective evaluation of our model to assess the perception of the synthesized gaze of a virtual character. While the perception of the gaze behavior of virtual characters has already been studied regarding the feeling of trust [30], the precise location of the gaze [31], or the subjective impressions depending on blink rates for human and cartoon-style characters [32], we decided to focus our study on the perceived realism of the gaze behavior and the awareness of the character. For this purpose, we conducted an online survey where participants were asked to judge the gaze of virtual characters animated with different configurations, according to different criteria.

5.1 Experiment Design

Task: In this online study, participants were presented with a number of 15-second videos displaying two virtual characters in the *lobby* scene used previously. The two characters are sitting at a table opposite of each other, with several objects positioned between them (Figure 4). The characters were designed to freely observe the virtual room, without any particular task. One of the character was facing the participant, while the other was facing away.

To evaluate the effects of the different components of our gaze model on visual realism and character awareness, participants were presented with eye-gaze animations generated using the same four configurations than in the ablation study presented previously (*i.e. Random*, *Saccadic*, *Saliency*, and *Complete*) as well as with eye-gaze animation driven from real recordings (*Actor*). In the *Actor* configuration, the character's eyes, head and eye blinks were animated using the recorded data of three participants randomly chosen from the user study presented in Section 4.1. We chose 15 seconds of the collected data corresponding to the beginning of their immersion when they discovered the *lobby*.

Participants were instructed to watch each video as many times as they wanted, before and while answering how much they agreed with the five following statements (answered on a 5-point Likert-scale ranging from 1: strongly disagree to 5: strongly agree), which were designed to judge the realism of eye-head animations:

- **S1** The character's eye activity is realistic, it resembles that of a real human
- **S2** The character seems aware of his environment, he is part of the scene
- **S3** Apart from what he is looking at, the movement of his eyes and head resemble those of a human
- **S4** The character seems to be looking at the opposite character or at the objects on the table
- **S5** Apart from what he is looking at, the movement of his eyes and head seem realistic to me

S_1 is intended to provide a general idea of the realism of the animation of the eyes, while S_3 and S_5 are more focused on the joint realism of the eye and head movements. We expect the *Random* and *Saliency* configurations to have low scores, as they do not account for any kinematics rule (no influence of the **saccadic map**). In contrast, S_2 and S_4 are related to the awareness of the character. We expect the *Random* and *Saccadic* configurations to have low scores, as

Fig. 4: Scene presented to participants in the subjective evaluation.

they do not take the environment into account (no influence of the **saliency map**).

Protocol: Upon starting the online experiment, participants were first asked to provide their age, gender, screen size, as well as experience in 3D animation. They were then presented in random order with 15 video clips (5 *Configurations* $\times$ 3 *Repetitions*), lasting 15 seconds each. For each *Configuration* condition, 3 different videos were rendered (*Repetitions*), either by running 3 different simulations or using sequences from three different real actors.

Participants: Fifty-three unpaid participants, recruited via internal mailing lists among students and staff, volunteered for this online experiment (37 F, 15 M, 1 NB; age: avg=30$\pm$10, min=21, max=63). They were all naive to the purpose of the experiment. Participants reported using screen sizes ranging from 14 to 34 inches.

5.2 Analysis

We asked participants how much they agreed with five statements, which we designed to provide information about the realism (S_1, S_3, S_5) and awareness (S_2, S_4) of the virtual character. To evaluate the degree of consistency between answers to the different statements, we first computed Cronbach's alphas. Overall, we found a high internal consistency between the statements related to realism ($\alpha > 0.88$), as well as between those related to awareness ($\alpha > 0.82$). In the rest of the analysis, we therefore only report results aggregated over the statements of each group, and that are significant at the 95% level (*i.e.*, $p < 0.05$). As the normality assumption was violated (Shapiro-Wilk test), we then performed Friedman tests to evaluate the effect of the different configurations on the perceived realism and awareness. Post-hoc comparisons were then performed using a Durbin Conover test. We therefore report the median values (M) when appropriate.

5.3 Results

First, as illustrated in Figure 5, results showed an effect of the gaze model on both realism ($\mathcal{X}^2(4) = 83.7, p < 0.001$) and awareness ($\mathcal{X}^2(4) = 141, p < 0.001$). Looking further into the effects, we found that the *Random* configuration was perceived as being less realistic ($M = 2.0$) than all the other configurations ($p < 0.001$), which were all rated high

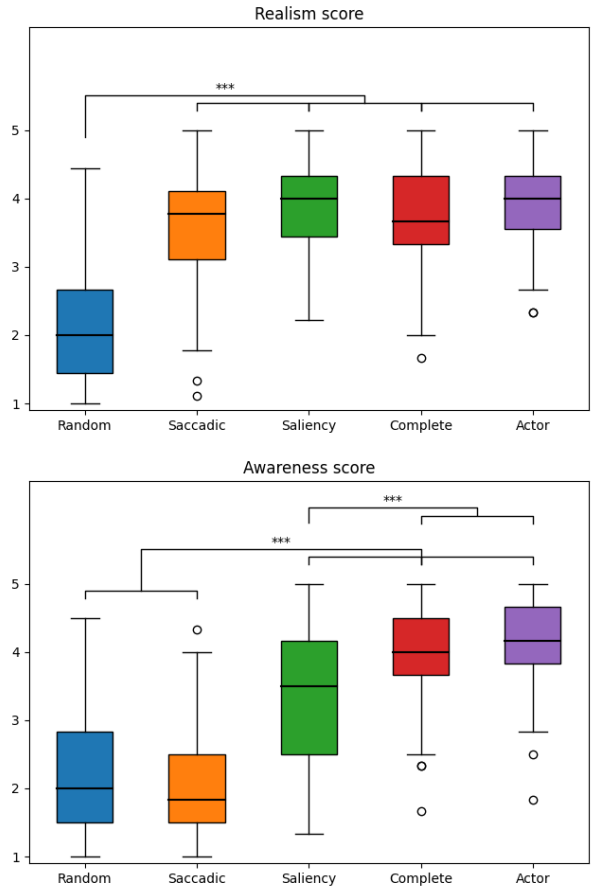


Fig. 5: Effect of different configurations of our model, as well as of real data from actors, on the perceived realism and awareness of the eye-head animations. The different configurations of our model (*Random*, *Saccadic*, *Saliency*, and *Complete*) are detailed in Section 4.2.

in terms of realism ($M > 3.7$). Interestingly, none of the *Saccadic* or *Saliency* configurations were rated significantly lower in realism than our *Complete* model or the *Actor*. This suggests that both saliency and saccadic information contribute to the final realism, but that their combination is not necessary to reach a higher level of realism. We however found slightly different results for awareness, where the *Random* and *Saccadic* configurations obtained the lowest ratings (resp. $M = 2.0$ and $M = 1.8$), then the *Saliency* configuration ($M = 3.5$), and finally both our *Complete* model and *Actor* (resp. $M = 4.0$ and $M = 4.2$). Unlike for realism, this suggests that saliency contributes more to the character's awareness than the saccadic information, but that their combination in our *Complete* model displays even higher levels of awareness, similar to real captured actors.

6 Discussion and Conclusion

Summary of contributions: In this paper, we presented a novel multi-map saliency-driven method simulating realistic gaze behaviors for non-conversational characters, which is usable in real-time interactive applications. Our model relies on three components (*saliency model*, *saccadic model* and *eye-head animation model*) that build on the lastest knowledge

and state-of-the-art results regarding eye-gaze activity, and which are combined for the first time in a unique real-time implementation which we make available to the community. It also enables users to easily customize novel features to synthesize specific gaze behaviors in a unified framework, unlike previous approaches which typically require to directly specify how to control the character's gaze through hand-crafted or scripted rules. This model provides a solid basis for the simulation of characters freely exploring their environment, as well as for generating more elaborate gaze behaviors. We demonstrated, through both objective and subjective evaluations, the achieved level of realism and the contributions of the various components of our solution. A key point in our approach is the coupling of a saliency and a saccadic model that can both be tuned by providing datasets corresponding to the desired context of usage. In the following, we will present the limitations of our approach.

Visual saliency model: We rely in our approach on the deep saliency model MSINet, which was trained to predict the saliency of natural images. Because of discrepancies between virtual and real scenes, its performance can be impacted when the visual rendering quality differs significantly from real images. In our implementation, we therefore use a physically-based rendering engine that realistically simulates light interactions to render High Dynamic Range images. Relying on a saliency model learned on natural images might however become more problematic in situations that involve low realism or non-photorealistic rendering, such as creating eye-gaze animations for cartoon characters, which might require to specifically train the saliency model using a different dataset (*e.g.*, comics [33]).

Saccadic model: Similarly, our saccadic model uses a single saccadic map for all our visual examples. However, previous work has shown that eye movements are specific to an environment and a task [34], and can also be influenced by the individual [35]. In an ideal situation, creating realistic eye-gaze animations would require to use a different saccadic map for each specific environment, task, character, etc. While this would be necessary to simulate specific behaviors, such as the gaze behavior of a person reading which displays most saccades to the right in an European-like culture and only occasionally to the left when starting a new line, creating these maps requires specific datasets which are extremely tedious to acquire. In specific cases, it would also be possible to rely on the existing literature on the subject, *e.g.*, to create saccadic maps typical of children behaviors [35], or representative of certain pathologies [36]. Furthermore, the saccadic map used in our model was abstracted from the data of the VR study for 360 videos [23]. This can potentially have an effect on the resulting realism, as VR is known to induce various biases, including eye-gaze differences compared to real situations [24], [37], whether it is due to the restricted field of view or to the absence of other stimuli (*e.g.* auditory). However, our results demonstrate that generic saccade maps as the one we used in our experiment already seem to provide high-realism results for our targeted application, *i.e.*, the generation of the gaze activity of characters freely exploring virtual scenes.

Animation model: The animation of the eyes and head of the virtual character is controlled based on velocity rules inferred from eye-tracking and head-tracking data. However, when observing closeup shots of the characters' eyes, successive saccades of small amplitudes can feel somewhat abrupt, with a lot of very fast visible eye movements, suggesting that the eye animation model could be improved in these cases. Nevertheless, the animation of the eyes and head is separated from predicting the next fixation in our model, and more accurate animation models accounting for these subtle effects could therefore be explored in the future.

Temporal dimension: One important limitation of our model is related to the attractiveness of moving elements in the visual field. Moving objects, or sudden changes in the field-of-view, are often associated with salient areas [38]. The saliency model we use processes each image independently, without accounting for the temporal aspect. Several authors proposed dynamic saliency models, designed to handle this temporal information (see the benchmark of Wang et al. [39] for more details), but still present a high computational cost that is not yet compatible with our real-time context. However, given the versatility of our framework, we believe that this aspect could be explored using a custom velocity map, or by adapting existing methods for moving objects [40] or animated meshes [41], which should compensate for the missing temporal aspect.

Scripting and customization: In addition, other behaviors that are not related to kinematic movements or saliency can be handled using customization maps. As mentioned in Section 3.3, those maps enable a full customizable and screenwriting behavior of gaze with ease of control. As we showed, it enables for instance characters to react to sudden events or to focus on objects with a particular semantic. We proposed a new way of scripting gaze behaviors by implementing simple maps and controlling their coefficient weights. However, finding a stable set of weights that can include any additional scripting map might prove difficult, and will probably require further optimization. Moreover, complex behaviors could also be modelled using these customization maps, *e.g.*, eye movements in the context of social interactions. For instance, to simulate a shy character one could design a map prohibiting fixations to occur on the eyes of other characters and more distributed in the lower area of their body, or get inspiration from the OCEAN personality model to generate variability as proposed for crowds [42]. We also plan to explore the benefits of designing other types of maps in the future, *e.g.*, based on other animation-driven semantics such as facial expressions, on optical flow information, on a sound-driven attention, etc.

Future work: The discussion above revealed some relevant directions for future work. Among them, we believe that usage in the context of Virtual Reality applications would raise new interesting questions related to presence and embodiment. Does being part of virtual worlds where characters can show, through gaze activity, an unprecedented level of awareness significantly improve the perceived level of presence? Does being gazed at with such new levels of realism change the perception of the virtual self? The gaze is such a key emotional and cognitive concern in these social setups that this appears to be the best direction to take for the future of our gaze animation approach.

7 Acknowledgments

This work was sponsored by the French National Research Agency under the Investments for the Future program (PIA) grant (ANR-21-ESRE-0030 CONTINUUM).

References

[1] K. Ruhland, S. Andrist, J. Badler, C. Peters, N. Badler, M. Gleicher, B. Mutlu, and R. Mcdonnell, "Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems," in *Eurographics 2014-State of the Art Reports*, 2014, pp. 69–91.

[2] M. F. P. Gillies and N. A. Dodgson, "Eye movements and attention for behavioural animation," *The Journal of Visualization and Computer Animation*, vol. 13, no. 5, pp. 287–300, 2002.

[3] O. Oyekoya, W. Steptoe, and A. Steed, "A saliency-based method of simulating visual attention in virtual scenes," in *Proceedings of the ACM Symp. on Virtual Reality Soft. and Tech.*, 2009, pp. 199–206.

[4] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*, vol. 5200, 2003, pp. 64–78.

[5] S. C. Khullar and N. I. Badler, "Where to look? automating attending behaviors of virtual human characters," *Autonomous Agents and Multi-Agent Systems*, vol. 4, no. 1, pp. 9–23, 2001.

[6] C. Peters, G. Castellano, M. Rehm, E. André, A. Raouzaiou, K. Rapantzikos, K. Karpouzis, G. Volpe, A. Camurri, and A. Vasalou, "Fundamentals of agent perception and attention modelling," in *Emotion-Oriented Systems*, 2011, pp. 293–319.

[7] C. Peters and C. O'Sullivan, "Bottom-up visual attention for virtual human animation," in *Proceedings 11th IEEE International Workshop on Program Comprehension*, 2003, pp. 111–117.

[8] N. Courty, E. Marchand, and B. Arnaldi, "A new application for saliency maps: Synthetic vision of autonomous actors," in *Proceedings of the International Conference on Image Processing*, vol. 3, 2003, pp. III–1065.

[9] E. Gu, J. Wang, and N. I. Badler, "Generating sequence of eye fixations using decision-theoretic attention model," in *International Workshop on Attention in Cognitive Systems*, 2005, pp. 277–292.

[10] A. Picot, G. Bailly, F. Elisei, and S. Raidt, "Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent," in *Int. Workshop on Intelligent Virtual Agents*, 2007, pp. 272–282.

[11] M. Gillies, "Practical behavioural animation based on vision and attention," CUCL, Tech. Rep., 2001.

[12] U. Ağıl and U. Güdükbay, "A group-based approach for gaze behavior of virtual crowds incorporating personalities," *Computer Animation and Virtual Worlds*, vol. 29, no. 5, p. e1806, 2018.

[13] C. Peters and C. O'Sullivan, "Attention-driven eye gaze and blinking for virtual humans," in *ACM SIGGRAPH 2003 Sketches & Applications*, 2003, pp. 1–1.

[14] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[15] S. P. Lee, J. B. Badler, and N. I. Badler, "Eyes alive," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, 2002, pp. 637–644.

[16] C. Pelachaud and M. Bilvi, "Modelling gaze behavior for conversational agents," in *International Workshop on Intelligent Virtual Agents*, 2003, pp. 93–100.

[17] R. Satogata, M. Kimoto, S. Yoshioka, M. Osawa, K. Shinozawa, and M. Imai, "Emergence of agent gaze behavior using interactive kinetics-based gaze direction model," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 433–435.

[18] S. Andrist, T. Pejsa, B. Mutlu, and M. Gleicher, "Designing effective gaze mechanisms for virtual agents," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2012, pp. 705–714.

[19] W. Steptoe, O. Oyekoya, and A. Steed, "Eyelid kinematics for virtual characters," *Computer animation and virtual worlds*, vol. 21, no. 3-4, pp. 161–171, 2010.

[20] C. Peters and A. Qureshi, "A head movement propensity model for animating gaze shifts and blinks of virtual characters," *Computers & Graphics*, vol. 34, no. 6, pp. 677–687, 2010.

[21] A. Klein, Z. Yumak, A. Beij, and A. F. van der Stappen, "Data-driven gaze animation using recurrent neural networks," in *Motion, Interaction and Games*, 2019, pp. 1–11.

[22] M. Abid, M. P. Da Silva, and P. L. Callet, "On the usage of visual saliency models for computer generated objects," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–5.

[23] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, p. 432–437.

[24] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in vr: How do people explore virtual environments?" *IEEE transactions on visualization and computer graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

[25] Y. Imaoka, A. Flury, and E. D. de Bruin, "Assessing saccadic eye movements with head-mounted display virtual reality technology," *Frontiers in Psychiatry*, vol. 11, p. 922, 2020.

[26] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[27] H. Jarodzka, K. Holmqvist, and M. Nyström, "A vector-based, multidimensional scanpath similarity measure," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 2010, p. 211–218.

[28] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT CSAIL, Tech. Rep., 2012.

[29] Y. A. D. Djilali, K. McGuinness, and N. E. O'Connor, "Simple baselines can fool 360° saliency metrics," in *IEEE/CVF International Conference on Computer Vision Workshops*, 2021, pp. 3743–3749.

[30] A. Normoyle, J. B. Badler, T. Fan, N. I. Badler, V. J. Cassol, and S. R. Musse, "Evaluating perceived trust from procedurally animated gaze," in *Proceedings of Motion on Games*, 2013, pp. 141–148.

[31] S. Loth, G. Horstmann, C. Osterbrink, and S. Kopp, "Accuracy of perceiving precisely gazing virtual agents," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018, pp. 263–268.

[32] K. Takashima, Y. Omori, Y. Yoshimoto, Y. Itoh, Y. Kitamura, and F. Kishino, "Effects of avatar's blinking animation on person impressions." in *Graphics Interface*, 2008, pp. 169–176.

[33] K. Bannier, E. Jain, and O. L. Meur, "Deepcomics: Saliency estimation for comics," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–5.

[34] M. Mills, A. Hollingworth, S. Van der Stigchel, L. Hoffman, and M. D. Dodd, "Examining the influence of task set on eye movements and fixations," *Journal of Vision*, vol. 11, no. 8, pp. 17–17, 2011.

[35] O. Le Meur, A. Coutrot, Z. Liu, P. Rämä, A. Le Roch, and A. Helo, "Your gaze betrays your age," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1892–1896.

[36] K. Bötzel, K. Rottach, and U. Büttner, "Normal and pathological saccadic dysmetria," *Brain*, vol. 116, no. 2, pp. 337–353, 04 1993.

[37] F. Berton, L. Hoyet, A.-H. Olivier, J. Bruneau, O. Le Meur, and J. Pettré, "Eye-gaze activity in crowds: impact of virtual reality and density," in *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2020, pp. 322–331.

[38] J. R. Brockmole and J. M. Henderson, "Object appearance, disappearance, and attention prioritization in real-world scenes," *Psychonomic Bulletin & Review*, vol. 12, pp. 1061–1067, 2005.

[39] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[40] S. Arpa, A. Bulbul, and T. Capin, "A decision theoretic approach to motion saliency in computer animations," in *International Conference on Motion in Games*, 2011, pp. 168–179.

[41] A. Bulbul, C. Koca, T. Capin, and U. Güdükbay, "Saliency for animated meshes with material properties," in *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, 2010, pp. 81–88.

[42] F. Durupinar, N. Pelechano, J. Allbeck, U. Güdükbay, and N. I. Badler, "How the ocean personality model affects the perception of crowds," *IEEE Computer Graphics and Applications*, vol. 31, no. 3, pp. 22–31, 2009.

[43] L. Sidenmark and H. Gellersen, "Eye, head and torso coordination during gaze shifts in virtual reality," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 27, no. 1, pp. 1–40, 2019.

[44] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, 2020.
[45] O. Le Meur and Z. Liu, "Saccadic model of eye movements for free-viewing condition," *Vision Research*, vol. 116, pp. 152–164, 2015.
[46] A. G. Samuel and D. Kat, "Inhibition of return: A graphical meta-analysis of its time course and an empirical test of its temporal and spatial properties," *Psychonomic bulletin & review*, vol. 10, no. 4, pp. 897–906, 2003.
[47] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1072–1080.
[48] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 679–700, 2021.
[49] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Computer Vision – ECCV 2018*, 2018, pp. 798–814.
[50] A. Bruckert, H. R. Tavakoli, Z. Liu, M. Christie, and O. Le Meur, "Deep saliency models: The quest for the loss function," *Neurocomputing*, vol. 453, pp. 693–704, 2021.
[51] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The international journal of aviation psychology*, vol. 3, no. 3, pp. 203–220, 1993.
[52] B. Keshavarz and H. Hecht, "Validating an efficient method to quantify motion sickness," *Human factors*, vol. 53, no. 4, pp. 415–426, 2011.
[53] M. Usoh, E. Catena, S. Arman, and M. Slater, "Using presence questionnaires in reality," *Presence*, vol. 9, no. 5, pp. 497–503, 2000.
[54] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, 2000, p. 71–78.
[55] B. R. Manor and E. Gordon, "Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks," *Journal of Neuroscience Methods*, vol. 128, no. 1, pp. 85–93, 2003.
[56] S. Stellmach, L. Nacke, and R. Dachselt, "Advanced gaze visualizations for three-dimensional virtual environments," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 2010, p. 109–112.

Ific Goudé received his PhD degree from the University of Rennes 1, France, in 2021. He is currently a researcher with Inria in the Virtus Team in Rennes, France. His research interests include real-time rendering, visual perception, machine learning and animation.

Alexandre Bruckert received his PhD degree from the University of Rennes 1, France, in 2022. He is a postdoctoral fellow at Nantes Université, in the LS2N lab. His research interests include human visual attention, computational models for simulating the gaze deployment, and visual perception-based applications.

Anne-Hélène Olivier received her PhD degree in sports sciences, from the University of Rennes 2, France, in 2008. She is currently an associate professor with the University of Rennes 2 (M2S and Virtus teams). Her research interests include the analysis and modeling of human walking and interactions between people. Her experimental approach uses virtual reality to investigate interactions between walkers, such as interactions between an individual and a crowd.

Julien Pettré received his PhD degree from the University of Toulouse III, in 2003, and Habilitation from the University of Rennes I, in 2015. He is senior researcher with Inria in the Virtus team. From 2004 to 2006, he was postdoctoral fellow at EPFL in Switzerland. He joined Inria, in 2006. His research interests include crowd simulation, computer animation, virtual reality, robot navigation and motion planning.

Rémi Cozot received his PhD and Habilitation degrees to conduct researches, in 1996 and 2014, respectively. He is professor at Université du Littoral Côte d'Opal. His main research topic concerns the preservation of a user's intent when editing real or computer generated contents. This topic involves many scientific skills such as visual human perception, image and video analysis, computer graphics. He has participated in many French and European projects on HDR imaging.

Kadi Bouatouch received his PhD degree from University of Nancy 1, in 1977 and a higher doctorate on computer science in the field of computer graphics in 1989 (University of Rennes 1). He is an electronics and automatic systems engineer (ENSEM 1974). He is currently Emeritus Professor at the University of Rennes 1 (France). His research interests cover global illumination, lighting simulation for complex environments, GPU based rendering and computer vision.

Marc Christie received his PhD degree from the University of Nantes, France, in 2003. He is currently an associate professor at University of Rennes 1. His focus is on virtual cinematography, which is the application of real cinematography techniques to virtual 3D environments.

Ludovic Hoyet received his PhD degree from INSA Rennes in 2010, and Habilitation from the University of Rennes 1 in 2022. He is a researcher with Inria in the Virtus team. He worked as a research fellow in Trinity College Dublin under the supervision of Pr. Carol O'Sullivan. His research interests include the realtime animation of virtual characters based on perceptual features.