Realistic and Interactive Robot Gaze

Matthew K.X.J. Pan*, Sungjoon Choi*, James Kennedy*, Kyna McIntosh*, Daniel Campos Zamora*,
Günter Niemeyer†, Joohyung Kim‡, Alexis Wieland §, and David Christensen*
*Disney Research, Glendale, California, USA
†California Institute of Technology, Pasadena, California, USA
‡University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
§Walt Disney Imagineering - Advanced Development, Glendale, California, USA

***Abstract*— This paper describes the development of a system for *lifelike* gaze in human-robot interactions using a humanoid *Audio-Animatronics*® bust. Previous work examining mutual gaze between robots and humans has focused on technical implementation. We present a general architecture that seeks not only to create gaze interactions from a technological standpoint, but also through the lens of character animation where the fidelity and believability of motion is paramount; that is, we seek to create an interaction which demonstrates the illusion of life. A complete system is described that perceives persons in the environment, identifies persons-of-interest based on salient actions, selects an appropriate gaze behavior, and executes high fidelity motions to respond to the stimuli. We use mechanisms that mimic motor and attention behaviors analogous to those observed in biological systems including attention habituation, saccades, and differences in motion bandwidth for actuators. Additionally, a subsumption architecture allows layering of simple motor movements to create increasingly complex behaviors which are able to interactively and realistically react to salient stimuli in the environment through subsuming lower levels of behavior. The result of this system is an interactive human-robot experience capable of human-like gaze behaviors.**

I. INTRODUCTION

Animatronic figures, more commonly known as animatronics, combine robotics with audio and visual elements to create a life-like character. Disney proprietary animatronics – referred to as *Audio-Animatronics*® figures – are extensively used in Disney theme parks to create repeatable shows and storytelling involving animal or human characters, providing consistent entertainment for guests. To achieve human-realistic behaviors on these types of robots, animators must manually design and refine trajectories to create fluid motions. Although preventing these behaviors from descending into the uncanny valley, this limits humanoid animatronics to performing non-interactive, scripted shows.

As off-the-shelf robotic sensing and actuation capabilities improve, there are increased opportunities for animatronic platforms to create more entertainment and engagement. Such animatronics can directly engage and interact with people to produce deeper immersion in storytelling. To advance the state-of-the-art of these robot figures, we aim to create interactive experiences that are human-like, dynamic, engaging and responsive.

To produce immersive and authentic shows, animatronic systems are required to create the *illusion of life*. In pursuit of this goal, we focus here on a single social cue: eye gaze.

Fig. 1. *Audio-Animatronics*® bust used to demonstrate the gaze system.

Gaze has been shown to be a key social signal, shaping perceptions of interaction partners. For example, people who make more eye contact with us are perceived to be similar to us, as well as more intelligent, conscientious, sincere, and trustworthy [1], [2], [3]. Furthermore, gaze appears to also convey complex social and emotional states [4], [5].

Given the importance of gaze in social interactions as well as its ability to communicate states and shape perceptions, it is apparent that gaze can function as a significant tool for an interactive robot character. Thus, the aim of this work is to develop a system to emulate human-like mutual gaze. In this paper, we present an architecture that can be used to

provide a robotic platform, shown in Fig. 1, with believable human-like gaze behaviors when interacting with groups of people.

This work seeks to combine the technical implementation of robot gaze with aspects of *animation* and *show* to create an illusion of life. We have specifically chosen to design robot gaze behavior using this approach rather than one grounded in human psychomotor behavior as we posit that animation offers a simpler framework for creating robot characters: it aims to focus on the appearance of behaviors rather than attempting to mimic the complex underlying mechanisms of human actions. Additionally, animation is more forgiving in that it allows for exaggerations of human qualities to create memorable and enjoyable characters.

As such, much of this work draws upon principles used by animators to bring cartoon characters to life [6], including the usage of 'arcs' (natural motion following an arched trajectory), 'overlapping action' and 'drag' (tendency for different body parts to move at different timings or frequencies), and 'slow in and slow out' (acceleration and deceleration at the beginning and end of actions, respectively). Such traditional animation techniques form the foundation for our method of inducing realism for our robot character.[1]

II. PRIOR WORK

The introduction outlined the importance of eye gaze for interactions between humans. These social functions of eye gaze also transfer to interactions between humans and robots. It has been established that people accurately perceive where robots are gazing [7]. Additionally, human-like gaze cues can be used to improve the perceived quality of interactions, such as in human-robot handovers [8].

Given the importance of eye gaze for interactions, researchers have focused not only on potential psychological effects that robot gaze cues might have on humans, but also on automated methods for generating believable eye gaze. One early example of such work comes from Breazeal and Scassellati [9]. Their attention mechanism relied on face, color, and motion detection, with an additional habituation input. These inputs were weighted by the motivations and emotions of their robot, ultimately resulting in a decision about whether to look at a person or a toy. The work that we present here shares the bottom-up approach of combining lower-level percepts to influence higher-level behaviors. More recently, Zaraki et al. implemented a context-dependent social gaze-control system implemented as part of a humanoid social robot [10]. The work presented in this paper also has some parallels with their research - particularly with relation to the use of proxemics, habituation and attention modeling.

Other prior work has used visual saliency, modeled after the way that human vision works [11], [12]. Vijayakumar et al. developed a system that uses a winner-take-all network to produce a saliency map from visual features [12]. However this presents challenges in ignoring robot self-motion, as well as stabilizing camera inputs as the robot moves. We avoid these issues by having a fixed location for the camera and utilizing existing knowledge of a visual scene (such as the presence of people or objects) to generate higher-level and more intentional behaviors.

Handcrafted animation of robots has previously been used to create instances of behaviors that can portray aspects of a character, such as personality [13]. However, the work here seeks to go into greater detail for a single social cue and provide more description about how a general architecture can realize human-like animation. An alternative to using handcrafted animation is to derive robot behavior from data. Deep learning approaches are showing promise in extracting behaviors from human interaction, e.g., [14], [15]. However, there is either a tendency toward a typically generic mean when large multi-person datasets are used, or a burden on the amount of data required from a single person. Issues with the adaptability of the output when in dynamic environments also present challenges. For example, data from dyadic interactions may not apply to triadic interactions. Given these challenges, we instead utilize handcrafted animation, where an animator has the control to express a character, supported by an architecture that enables dynamic adaptation for interactions.

III. PLATFORM AND SHOW

A. Robot Platform

The robot platform used in this work is a custom Walt Disney Imagineering *Audio-Animatronics®* bust consisting of head and upper torso, as shown in Fig. 1. The figure has 19 degrees-of-freedom (DOF). Of these DOFs, we only use nine: the neck (3), eyes (2), eyelids (2), and eyebrows (2). The remaining DOFs produce jaw and lip movement and are not used in the current system. The robot is controlled via a custom proprietary software stack operating on a 100Hz real-time loop.

B. Perception

A Mynt Eye D-1000 RGB-D camera[2] is used for perception of people within the robot's field-of-view (FOV). The camera has a horizontal and vertical FOV of 105° and 58°, respectively, and an approximate range of 0.3 to 10 meters. This is more limiting than the typical human FOV (H: 200° V: 135°) and range. The camera is mounted to the upper torso of the robot's figure and is stationary. Skeleton fitting is performed via a perception engine which uses wrnch[3] at its core. The perception engine fits skeletons in 2D and then depth-projects to provide 3D joint locations. Each skeleton is composed of the following points of interest: eyes, nose, ears, neck, shoulders, elbows, wrist, hips, knees, ankles, big toes, small toes and heels. Points-of-interest are transformed from the camera frame to the robot's frame of reference.

[1]This work describes a robot system that is intended to create visually appealing movements. It is recommended to view the video supplementing this manuscript to gain a better understanding of the contribution of the work.

[2]https://mynteye.com/

[3]https://wrnch.ai/

C. Show Context

The core purpose of animatronics figures is to deliver a *show*, in the same manner that actors perform in a theatre. Just like actors, animatronics take on personas and characters and are given scripts and blocking to tell a story and/or enact a scene. With this in mind, we created a show context and persona for our robot character to better frame the concepts we wished to develop in the robot gaze system.

The robot character plays an elderly man reading a book, perhaps in a library or on a park bench. He has difficulty hearing and his eyesight is in decline. Even so, he is constantly distracted from reading by people passing by or coming up to greet him. Most times, he glances at people moving quickly in the distance, but as people encroach into his personal space, he will stare with disapproval for the interruption, or provide those that are familiar to him with friendly acknowledgment.

With this context and the persona of an elderly man, we are able to test behaviors such as glancing and mutual gaze in a realistic scenario while catering to some limitations of our hardware, i.e., limited FOV of camera and system latency. For this scenario, we specifically chose not to provide our robot with audition as we wanted to focus on understanding eye gaze in a scenario that does not depend on conversational content.

IV. ARCHITECTURE

The architecture of our robot gaze interaction consists of three components: the attention engine, behavior selection engine, and behavior library (Fig. 2). These components all have bi-directional communication with one another and ultimately drive animations on the robot character.

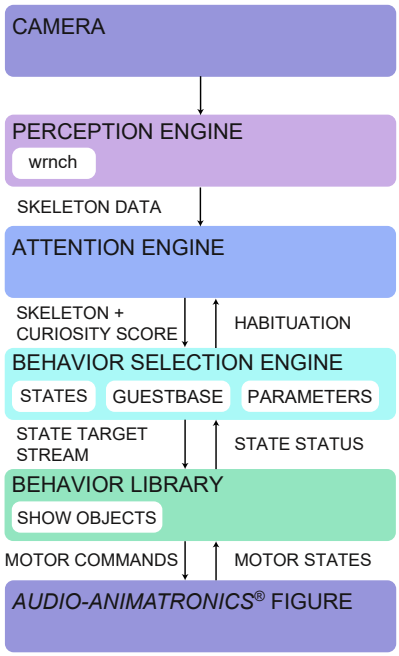

Fig. 2. Overall system architecture.

A. Attention Engine

In a similar fashion to prior work such as [9], the attention engine identifies lower-level salient stimuli in the environment. Currently, the system only examines stimuli generated by persons in the scene based on movements of the fitted skeletons. The attention engine generates a 'curiosity score' assigned to that person indicating their salience/significance as well as how important it is for the robot character to respond to them.

For this approach, the attention engine consumes 3D positions of all people in the visual scene from the camera data and smooths the incoming data over time. A numerical value is assigned to each detected skeleton to differentiate persons in the scene. Using this data, the attention engine provides estimates for whether people are performing certain actions, such as waving, as well as calculating how quickly people are moving over time. The Euclidean distance to the robot is used as an additional proxemic factor.

For each unique person identified in the scene, the curiosity score Φ is computed based on features which are chosen to be of importance. It is meant to be a customizable measure of salience and is calculated using a function in the form of Eqn. (1).

$$\begin{aligned}\Phi = \Theta(t) * w_{distance} * f(d_{nose}) \\ * \big[w_{raised\_hand}(h_{r\_hand} + h_{l\_hand}) \\ + w_{hand}(\hat{v}_{r\_hand} + \hat{v}_{l\_hand}) \\ + w_{nose}\hat{v}_{nose} + w_{distance}f(d_{nose})\big] \quad (1)\end{aligned}$$

where $\Theta(t) \in [0, 1]$ is a habituation factor at time t (described in more detail below), w_x is a weighting for factor x, $\hat{v}_x$ is the velocity of x normalized over an average maximum velocity detected of that feature, $h_{r\_hand}, h_{l\_hand} \in \mathbb{B}$ and represent whether the hands are raised above the shoulder, and $f(d) = e^{-ad_{nose}}$ representing an exponential decay with rate a over Euclidean distance d_{nose} (where d_{nose} is in meters) from the person's nose to the robot.

As seen from Eqn. (1), the way in which curiosity is calculated in our current system is fairly simplistic, being based upon locations and velocities of the hands and nose. This selection of features was loosely informed by the way people attempt to gain attention from another person - e.g., moving closer/quickly to the person and waving their hands. Although this example curiosity function uses kinematic features of detected people in the scene, it can be expanded to include other salient environmental features. Such features could include specific clothing colors, auditory cues, facial emotions, props, etc. The weights of the function are customizable and can be tuned to be sensitive to some features and not others. This design enables character creators to shape the interactive personality of the character through modification of weights. As future work, we hope to develop a method to automatically tune these weights such that they fit attention behaviors of characters we wish to implement on the platform.

Another function of the attention engine is to receive feedback from the behavior selection engine (see Sec. IV-B on which persons-of-interest have been attended to. Persons who have been attended to by the system via mutual gaze actions have their Φ decreased linearly over time. In a biological system, this is similar to a process known as habituation to stimuli and has been included in other robotic gaze systems [9], [16]. Our habituation function limits the character from repeatedly responding to a single guest and ignoring others. The function also includes a mechanism to restore the Φ of persons if they are not being attended to/gazed upon by the robot. This allows persons who have been gazed at by the robot character and subjected to habituation to regain the ability to catch the robot's attention after a period of time. This behavior is modeled by Eqn. (2):

$$\Theta(t+\Delta t) = \Theta(t) + (\Gamma * m_{hab} + \bar{\Gamma} * m_{rest})\Delta t \cup [0,1] \quad (2)$$

where $\Theta(t+\Delta t) \in [0,1]$ is the habituation value at $t+\Delta t$, $\Gamma \in \mathbb{B}$ is equal to 1 when the robot is looking at the person and 0 otherwise, m_{hab} is a negative habituation decay rate, and m_{rest} is a positive restoration rate.

B. Behavior Selection Engine

The behavior selection engine is the component of the robot character system that represents higher-level reasoning for the character. This component contains a handcrafted state machine that directs which behavioral state the robot is in. The behavior selection engine also maintains information regarding parameters of the robot (e.g., current state, curiosity thresholds, state timeout durations) as well as skeletons which are seen by the camera.

1) States: Given the context of the system, four states were devised to cover the aspects of behavior required to deliver the interactive experience. The behavior of each state is described in the subsequent subsection (Sec. IV-C). Here we present the states (depicted in Fig. 3) and how their transitions are determined:

- **Read**: The Read state can be considered the 'default' state of the character. When not executing another state, the robot character will return to the Read state. Here, the character will appear to read a book located at torso level.
- **Glance**: A transition to the Glance state from the Read or Engage states occurs when the attention engine indicates that there is a stimuli with a curiosity score Φ above a certain threshold, but below the threshold that would trigger the Engage state. In this state, the character gazes at the disruptive person (which we refer to as the person-of-interest) with his eyes and a varying degree of head motion away from its previous state (reading or engaging) based on Φ. This state is exited after a timeout (also a function of Φ) and returns to the Read state, or when another guest has a high enough Φ to trigger the Engage state.
- **Engage**: The Engage state occurs when the attention engine indicates that there is a stimuli with large enough Φ to meet a threshold and can be triggered from both Read and Glance states. This state causes the robot to gaze at the person-of-interest with both the eyes and head. The state is exited after both of the following are true: the time in the engage state meets a minimum timeout duration and when Φ of the person-of-interest decreases below the Engage threshold; the timeout guarantees a minimum time length for this behavior to ensure character believability and naturalness.
- **Acknowledge**: The Acknowledge state is triggered from either Engage or Glance states when the person-of-interest is deemed to be familiar to the robot. Currently, familiarity is a property that is randomly assigned to detected persons.

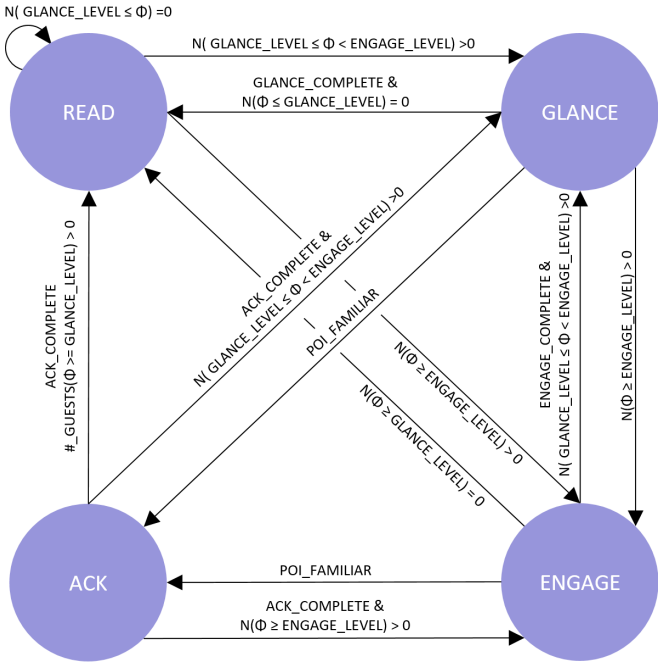

Fig. 3. States in the behavior selection engine. N(condition) represents number of detected people in the scene who meet the condition.

2) Guestbase: In addition to a state machine, the behavior selection engine also contains a database of people in the scene, called the guestbase, which functions as a form of short-term memory. Each entry in the guestbase contains the numerical value associated with each guest's skeleton, times that guest was first and last seen, 3D skeleton data, current curiosity score Φ, whether the guest has previously been a person-of-interest, and the character's familiarity with that guest. Persons are added to the guestbase as they appear in the scene, and are removed after they leave the scene.

The purpose of the guestbase is threefold: 1) to store 3D points of persons – in particular the locations of eyes and nose – such that the robot character can attempt mutual gaze if the guest becomes a person-of-interest, 2) to keep a record of those people that have been persons-of-interest for the purpose of habituation (Sec. IV-A), and 3) to denote special properties of that guest with relation to the system, e.g., whether the guest is familiar to the character (Sec. IV-B.1).

C. Behavior Library

The behavior library consists of show animation definitions that can be used to drive motor actions and movements on the *Audio-Animatronics*® figure.

1) Show Objects: The library is constructed using a concept of *show objects*, where each show object consists of a set of motor commands for the *Audio-Animatronics*® figure. Show objects can be *grouped* (made to run simultaneously), *sequenced* (run in order one after another), *looped* (repeated for a duration), and *randomly selected* (pick one show object randomly from a set). Each operation can produce a new, singular, show object. Additionally, show objects can be added together, multiplied by scalars, or subtracted from each other. Thus, show objects provide a powerful interface for constructing complex motor behaviors using simpler building blocks. As an example of this concept in action, the reading animation pseudocode is shown below as Algorithm 1.

Algorithm 1 Example Reading Show Pseudocode

```
1:  function READ
                                          ▷ Define initial poses
2:      head_down ← move_to (
            -0.33rad for 1.0s with head_ud)
3:      eyes_down ← move_to (
            -0.40rad for 1.0s with eyes_ud)
4:      eyes_left ← move_to (
            0.20rad for 1.0s with eyes_lr)
5:      lids_down ← move_to (
            70% for 1.0s with eyelids)
                                          ▷ Group poses together
6:      read_pose ← group (
            head_down, eyes_down, eyes_left, lids_down)
                ▷ Create reading actions using keyframes
7:      head_read ← keyframes (
            [0.20rad@t=0s,
            -0.20rad@t=3.8s,
            0.20rad@t=4.6s] with head_lr)
8:      eyes_read_A ← keyframes (keyframes using eyes_lr)
9:      eyes_read_B ← keyframes (keyframes using eyes_lr)
10:     eyes_read_C ← keyframes (keyframes using eyes_lr)
                    ▷ Build up show objects from primitives
11:     eyes_read ← random_select (
            eyes_read_A, eyes_read_B, eyes_read_C)
12:     read_loop ← loop ( group( head_read, eyes_read))
13:     read_show ← sequence( read_pose, read_loop)
        return read_show
```

As shown in the reading example, show objects which contain simple poses and keyframe animations are combined with other show objects to create increasingly complex behaviors which eventually form a complete reading show.

2) Behavior Layering: Having fully-formed show elements, we require a method for organizing behaviors. We do so by loosely subscribing to a subsumption architecture: a method of robot behavior and control proposed by Rodney Brooks which layers behaviors hierarchically [17]. This layering system is organized by 'levels of competence', where lower levels represent more basic functions of the robot, and higher levels represent behaviors which require advanced processing. In this architecture, higher levels are able to 'subsume' lower levels – i.e., they are able to integrate, modify, suppress or even completely override lower levels. Our implementation uses the following levels of show objects (in increasing order of complexity):

- L0 **Zero Show**: This level is the most basic level of the robot's behavior. It sets the motors to a default 'zero' position for the show. For example, the jaw is set to be closed at this level (as opposed to hanging open). The jaw remains closed as no higher level of behavior subsumes its actuators.
- L1 **Alive Show**: This level provides the minimum requirements for the robot to be perceived as animate. This includes behaviors such as breathing, eye blinking, saccades, etc. Such items would fall under the animation principle of 'secondary action', i.e., secondary behaviors which add to the show but do not interfere with higher level behaviors [6].
- L2 **Read Show**: This is the primary behavior of the show corresponding to the Read state and is always running. It forms the default fallback behavior once higher levels of behavior give up control of the system.
- L3 **Glance Show**: The glance show represents a glancing behavior directed towards salient stimuli as perceived by the attention engine. A greater curiosity score Φ will increase the degree in which the head is tilted towards the stimuli and away from its position in the previous state.
- L4 **Engage Show**: The engage show represents the character's full attention being directed towards a stimuli, and subsumes both the read and glance show. The head and eyes are directed towards the stimuli. Normal eye saccading from the L1 alive show is interrupted to saccade based on a person's eye and nose locations as perceived by the camera. Eyebrows are furrowed to give the character some contempt for being distracted.
- L5 **Acknowledge Show**: The acknowledge show is run whenever people are recognized as being familiar and subsumes the engage show. It randomly selects a nodding action from a predefined set of keyframe show objects involving only the neck actuators and plays it as an overlay on top of the engage show. These nodding actions are based on actor motion capture data that was retargeted to this robot using the algorithm presented in [18].

D. Implementation Lessons

1) Saccades: Although saccades have been implemented on humanoid robot platforms in prior work [19], [10], [20], they often do not match the performance and behavior of human saccades during gaze. This may be in part due to motor speed limitations, however, many humanoid robot designs integrate cameras into the eyes or on the head, e.g., SoftBank Pepper[4], EMYS[5]. This design choice introduces a number of

[4]https://www.softbankrobotics.com/us/pepper
[5]https://emys.co/product

non-trivial challenges such as motion blur and distortion of camera images during eye movements. Thus, the eyes require carefully tuned controllers and/or slow, smooth eye motion profiles that lack fidelity when compared human eye motions.

Having a static camera system mounted externally avoids such issues and allows us to design saccades that better mimic those observed in humans. Our *Audio-Animatronics*® figure is able to saccade with a frequency of 20 Hz. For our mutual gaze actions (glance and engage), the eyes saccade randomly between three points forming a triangle on a person-of-interest: the eyes and nose. Timing of saccadic movements is random with a range of 0.1 to 0.5s.

2) Motion Bandwidth: To increase realism of movements on the *Audio-Animatronics*® figure, we leveraged the animation principle of *overlapping action*, where some parts of a character naturally move faster than others during motion [6]. For our bust, this meant ensuring that the servo rate of each DOF in the head and neck were similar to human movement. For example, eyelid motion had a frequency of 10 Hz, whereas head/neck motion was limited to 5 Hz or less. One particularity notable implementation of this principle is that with any turn-to-glance or turn-to-engage movement, the head turning is always led with the eyes, i.e., the eyes reach the gaze position first at a high rate of speed, followed by the head at a lower rate of speed, giving rise to a realistic transition into mutual gazing.

V. DISCUSSION

A. Believable Interactions

The architecture and implementation for human-like robot gaze appears to produce life-like gaze on our robot character. However, we posit that the believability of this implementation is a function of time and distance. At further distances, lower fidelity behaviors can seem believable, and the same is true for shorter interaction times. At closer distances and for longer periods of time, more complex behaviors are required to create believable characters. For example, simple head motion may be believable from a distance, but as individuals move closer to a robot, the illusion of life would break, and eye gaze becomes essential. The scenario we explore in this work is satisfactory at close distances, but for a relatively short period of time (one or two minutes).

The implementation described here produces a believable gaze behavior for this time scale while also providing behavioral building blocks for longer time scales, but other challenges emerge at this interaction distance. The physical appearance of the robot is a key concern as it relates to the uncanny valley, and is the subject of research [21]. Physical interaction is potentially expected in such proxemically-close interactions too, including such capabilities as handshakes [22] or handovers [8]. The interdependence of these behaviors and various social cues adds to the complexity of believable character animation in such a dynamic environment.

Our character displays some small variance of emotion when choosing whether to stare at or acknowledge a person. However, the lack of emotional range is a limitation in the current work. Previous work has explored how robots might be able to use social cues to display different emotions whilst still being perceived as a specified personality [23]. In future work, we aim to take this further and consider how robot motion should portray transitions in emotional state.

B. Tuning Attention

The attention engine presented in this work is an attempt at enabling a robot to identify, sort, rank, and habituate salient stimuli in its environment, much like others have done previously in [9], [11], [12] using raw camera data. However, because our method acts on more concrete observations of stimuli (e.g., skeletal kinematics) and is customizable, it offers those programming characters on animatronic platforms – i.e., animators – transparency and intuitive control over how saliency is perceived by a character. For example, an animator may wish to create a nervous character that responds to the slightest sound or movement, whereas a character falling asleep may respond only to the most arousing stimuli. Being able to offer a set of controls to tweak character attention is valuable.

As future work, however, we would like to explore how attention engine parameters can be tuned more abstractly using learning-based dimension reduction to generate appropriate attention profiles for a variety of characters. Such an approach could reduce dimensionality of the control palette, making it easier for animators to select desired character attentiveness and habituation.

C. Subsumption can Easily Create Complexity

Although originally intended for control of mobile robots, we find that the subsumption architecture, as presented in [17], lends itself as a framework for organizing animatronic behaviors. This is due to the analogous use of subsumption in human behavior: human psychomotor behavior can be intuitively modeled as as layered behaviors with incoming sensory inputs, where higher behavioral levels are able to subsume lower behaviors. At the lowest level, we have involuntary movements such as heartbeats, breathing and blinking. However, higher behavioral responses can take over and control lower level behaviors, e.g., fight-or-flight response can induce faster heart rate and breathing. As our robot character is modeled after human morphology, mimicking biological behaviors through the of use a bottom-up approach is straightforward.

Additionally, the layering approach allows for the generation of complex motor animation quickly and easily, as opposed to having to manually package all required behaviors within the trajectory for each motor. Changing something like how fast the robot saccades, blinks or breathes only involves changing a single behavioral layer, rather than requiring the regeneration of an entire set of motor trajectories.

D. Saccades Increase Realism

One notable observation from the development of our gaze system is that saccades appear to significantly improve character realism. Vergence of gaze is the ability for eyes to focus on objects both near and far. For this, each eye

is required to move independently, i.e., eyes rotate towards each other to look at close objects, and away from each other for distant objects. However, our *Audio-Animatronics*® platform is physically incapable of vergence as the eyes are mechanically linked with the focal plane set at infinity, i.e., the eyes have parallel sightlines. Thus, during mutual gaze, the robot appears to look *through* you instead of *at* you.

However, we found that eye saccades provide the illusion of vergence on our platform with mechanically-linked eye movement. We hypothesize that this is due to human inability to focus on both eyes at once during mutual gaze; instead, we focus on one eye and then the other. The movement of the robot's eyes during saccading targets the face of the person (eyes/nose), thereby making it difficult for people to determine the robot's true focal plane and creating the illusion that both eyes are focused on the person.

VI. CONCLUSION

This research adds to an active domain in non-verbal social human-robot interaction. Previous research has shown the importance of gaze for improving legibility and efficiency of interactions between robots and humans. We extend this in a complementary way to demonstrate how a gaze system can be constructed to interactively engage people such that the robot may feel alive and responsive. We have developed a system in which a human-like *Audio-Animatronics*® figure can use gaze in a manner that appears consistent with human behavior using techniques drawn from research on human/robotic gaze and animation.

Through layering of simple behaviors, it appears that we are able to generate complex responses to environmental stimuli. This architecture is highly extensible and can be used to create increasingly complex animatronic gaze behaviors as well as other interactive shows. We see this work as an attempt to ascend from the uncanny valley through layering of interactive kinematic behaviors and sensorimotor responsiveness.

REFERENCES

[1] H. Kreysa, L. Kessler, and S. R. Schweinberger, "Direct Speaker Gaze Promotes Trust in Truth-Ambiguous Statements," *PLOS ONE*, vol. 11, p. e0162291, Sep 2016.

[2] N. A. Murphy, J. A. Hall, and C. R. Colvin, "Accurate Intelligence Assessments in Social Interactions: Mediators and Gender Effects," *Journal of Personality*, vol. 71, pp. 465–493, Jun 2003.

[3] C. Zhou, T. Jiang, and L. Zhu, "Direct Gaze Blurs Self-Other Boundaries," *Journal of General Psychology*, vol. 145, pp. 280–295, Jul 2018.

[4] C. L. Kleinke and P. D. Pohlen, "Affective and emotional responses as a function of other person's gaze and cooperativeness in a two-person game," *Journal of Personality and Social Psychology*, vol. 17, no. 3, p. 308, 1971.

[5] A. J. Van Breemen, "Animation engine for believable interactive user-interface robots," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 2873–2878, 2004.

[6] O. Johnston and F. Thomas, *The illusion of life: Disney animation*. Disney Editions New York, 1981.

[7] M. Imai, T. Kanda, T. Ono, H. Ishiguro, and K. Mase, "Robot mediated round table: Analysis of the effect of robot's gaze," in *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication*, pp. 411–416, IEEE, 2002.

[8] A. Moon, D. M. Troniak, B. Gleeson, M. K. Pan, M. Zeng, B. A. Blumer, K. MacLean, and E. A. Croft, "Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction - HRI '14*, (Bielefeld, Germany), pp. 334–341, ACM Press, Mar 2014.

[9] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, pp. 1146–1151, 1999.

[10] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 157–168, 2014.

[11] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Transactions on Robotics*, vol. 25, no. 4, pp. 861–873, 2009.

[12] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 4, pp. 2332–2337, IEEE, 2001.

[13] P. Mota, M. Paetzel, A. Fox, A. Amini, S. Srinivasan, J. Kennedy, and J. F. Lehman, "Expressing coherent personality with incremental acquisition of multimodal behaviors," in *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2018.

[14] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104, 2019.

[15] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pp. 4303–4309, IEEE, 2019.

[16] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 31, no. 5, pp. 443–453, 2001.

[17] R. A. Brooks, "A Robust Layered Control System For A Mobile Robot," *IEEE Journal on Robotics and Automation*, vol. 2, no. 1, pp. 14–23, 1986.

[18] S. Choi and J. Kim, "Towards a Natural Motion Generator: a Pipeline to Control a Humanoid based on Motion Data," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4373–4380, IEEE, Jan 2020.

[19] A. Ude, V. Wyart, L. H. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots*, vol. 2005, pp. 381–386, 2005.

[20] A. Roncone, U. Pattacini, G. Metta, and L. Natale, "A Cartesian 6-DoF Gaze Controller for Humanoid Robots," in *Robotics: Science and Systems*, Jun 2016.

[21] S. Nishio, H. Ishiguro, and N. Hagita, "Geminoid: Teleoperated android of an existing person," *Humanoid robots: New developments*, vol. 14, pp. 343–352, 2007.

[22] D. Mura, E. Knoop, M. G. Catalano, G. Grioli, M. Bächer, and A. Bicchi, "On the role of stiffness and synchronization in human–robot handshaking," *The International Journal of Robotics Research*, p. 0278364920903792, 2020.

[23] M. Paetzel, J. Kennedy, G. Castellano, and J. F. Lehman, "Incremental acquisition and reuse of multimodal affective behaviors in a conversational agent," in *Proceedings of the 6th International Conference on Human-Agent Interaction (HAI)*, pp. 92–100, ACM, 2018.