Received: 16 March 2017 | Accepted: 18 March 2017
DOI: 10.1002/cav.1757

SPECIAL ISSUE PAPER

WILEY

Autonomous social gaze model for an interactive virtual character in real-life settings

Zerrin Yumak | Bram van den Brink | Arjan Egges

Utrecht University, Utrecht, Netherlands

Correspondence
Zerrin Yumak, Utrecht University, Utrecht, The Netherlands.
Email: z.yumak@uu.nl

Funding information
Horizon 2020 RAGE—Realizing an Applied Gaming Eco-system, Grant/Award Number: 644187; Utrecht University Game Research Seed Money

Abstract

This paper presents a gaze behavior model for an interactive virtual character situated in the real world. We are interested in estimating which user has an intention to interact, in other words which user is engaged with the virtual character. The model takes into account behavioral cues such as proximity, velocity, posture, and sound; estimates an engagement score; and drives the gaze behavior of the virtual character. Initially, we assign equal weights to these features. Using data collected in a real setting, we analyze which features have higher importance. We found that the model with weighted features correlates better with the ground-truth data.

KEYWORDS
engagement, gaze model, multiparty interactions, interactive virtual humans

1 | INTRODUCTION

Gaze movement is important for modeling realistic social interactions with virtual humans. Although gaze animation based on low-level kinematics is well studied, autonomous generation of gaze at the high-level during social interactions and in real settings still remains as a challenge.[1] One of the open problems is how to drive the gaze behavior of an interactive virtual character situated in a real environment, that is, a virtual receptionist. It requires understanding which user has an intention to interact with the virtual character, in other words which user is more engaged. The users might be approaching the virtual character alone, in groups or they might just be passing by.

Recognition of goals, intentions, and emotions of other people is important for a fluent communication. If one has to give humanlike capabilities to artificial characters, they should also be able to predict the intentions of others. In this paper, we focus on the engagement detection problem as a prerequisite to initiating a conversation with a user and propose a model to autonomously drive the gaze behavior of the virtual character. Figure 1 shows our Virtual Character Sara interacting with a group of users.

Previous work modeled engagement based on heuristic rules.[2,3] It has been shown that machine learning approaches[4,5] outperform the basic heuristics. Both approaches have advantages and disadvantages. Although the former does not involve extensive validations of their model, the latter depends on huge data collection and analysis efforts. An overview of multiparty interactions and a discussion on open research challenges can be found in our previous work.[6] In this paper, our contribution is twofold: (a) We present a practical and general engagement model combining multiple behavioral cues to drive the gaze of an interactive virtual character. (b) We find the importance of these behavioral cues based on data collected in a real environment.

In Section 2, we mention the related work and present our contributions. In Section 3, we explain our engagement-driven gaze model for a 3D virtual character. Section 4 describes an experiment we conducted using our system and provides an analysis and discussion of the results. Section 5 concludes the paper and points out the future work.

2 | RELATED WORK

The analysis of group interaction dynamics in human–human interactions has been subject to attention in the area of social psychology and non-verbal communication.[7] The study of engagement has been mainly done in these communities and

FIGURE 1 Interactive virtual human Sara

less in the field of human–computer interaction.[2] Developing a model of engagement or any other high-level social behavior is first based on understanding the human behavior. Then, the model can be developed based on insights derived from human–human interactions.

The term engagement was first coined by Sidner et al[2] during their experiment with the penguin robot Merl in Mitsubishi Electric Research Labs. They defined engagement as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another". Peters et al.[8] mentioned that engagement is often related to interest, which is an emotional state linked to the participant's goal of receiving and elaborating new and potentially useful knowledge. They defined engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participants and of continuing the interaction." According to Peter et al,[8] engagement is related to the two stages of interaction: (a) before the interaction—The participant decides whether it is worth to start the interaction or (b) during the interaction—to monitor the continuous level of engagement. Michalowski et al[3] developed a receptionist robot and defined engagement based on proximity. They defined four states: present, attending, engaged, and interacting. They observed that the robot greeted late (meaning earlier anticipation is needed) and greeted people who did not intend to engage (meaning more accurate anticipation is needed).

Recent research focused on estimating engagement using data-driven approaches.[4] It has been found that machine learning approaches outperform the basic heuristics due to two reasons: (a) Fusion of various features help to make more accurate decisions. For example, a user might be close to the artificial agent but his or her posture might show that he or she is not really interested in engaging. (b) Machine learning is based on large data sets covering various cases, and these options might not be covered well by the heuristic methods.

In Bohus and Horvitz,[4] Bohus et al. describe an engagement model that senses the engagement state of multiple users and that makes high-level engagement control decisions to decide whom to engage. They first applied a heuristic engagement estimation method assuming a person is engaged if there is a frontal face in front of the camera. Then, the moments before engagement is labeled automatically to train the system without hand labeling. They applied a maximum entropy model to detect engagement combining several features such as location of the face, width and height, confidence score of the face, trajectory of location features, and attention. They took into account both the domain-dependent and domain-independent aspects of engagement. However, they do not mention which features among others are more important for engagement.

In Foster et al,[5] a data-driven engagement estimation method is described in a bar set-up where the robot serves drinks to multiple customers. Engagement state of the user involves whether the user comes close to the bar and makes eye contact with the bartender. Using the first prototype of the system developed based on rules, the authors collected interaction data, which were later annotated for data-driven engagement estimation. They trained the system using a set of classifiers using face and head coordinates, orientation, and sound as features. They found that face, right hand, and sound had higher importance in the model. Xu et al[9] developed an engagement-aware agent in multiparty conversations using a data-driven approach. They detected both engagement and disengagement intentions of users using support vector machines based on features such as direction of attention, change of speaking status, change of emotions, and distance. They also do not report which features are more important in engagement detection, although they mention the important features for attention.

Engagement and attention are terms often used in a similar context. Although research work on attention and engagement are related, these two concepts are different. Although attention is bottom-up and biology-driven, engagement is top-down and focuses on high-level social situations. Attention can be on various things such as objects, people, and actions, that is, attention to a fast moving object passing on the street. Engagement is the willingness to socially interact. In other words, higher attention is given to the person who is trying to engage in a reciprocal social interaction. Zaraki et al,[10] Kokkinara et al,[11] Xu et al,[9] and Grillon et al,[12] developed attention models for virtual humans and robots using multiple features such as proximity and velocity.

In this paper, we present a practical and general engagement model to drive the gaze behavior of an interactive virtual human during group interactions. We take into account multiple features for the computational model of engagement and initially assign equal weights to these features. Using data collected in a real setting, we find which features have higher importance in our engagement model. To our knowledge, that is the first study to take into account multiple people interacting in a truly open space considering the importance of a wide range of features for engagement detection.

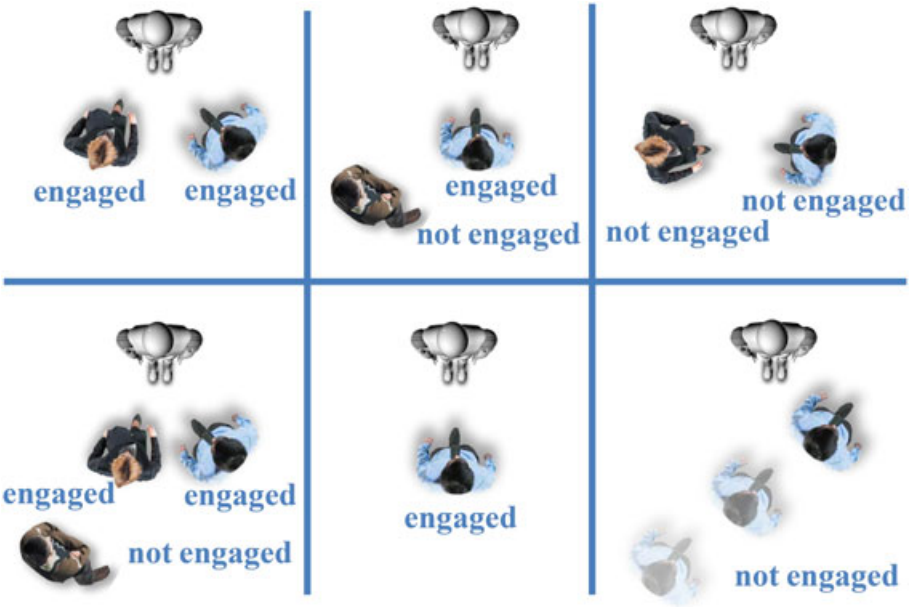

FIGURE 2 Examples of engaged and not-engaged situations

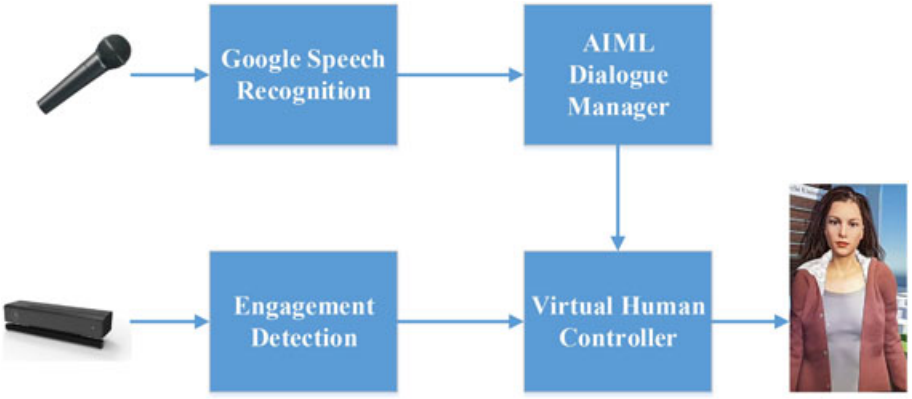

FIGURE 3 Overall architecture

3 | ENGAGEMENT-DRIVEN GAZE MODEL

People often group themselves into clusters, lines or circles, or various other kind of patterns.[7] Group members come together in a way that the space between them is shared to allow their equal contribution. It happens many times in daily life when people form a conversational group for casual talk. Participants stand so that they face inwards to a small space, which they cooperate together. People outside the circle are the bystanders or passer-by people, and they are not engaged. Figure 2 shows various interaction configurations with an artificial agent and the engagement state of the participants.

In this section, we explain our engagement-driven social gaze model. First, we describe the overall architecture of our system. Then, we describe our engagement model and the features that are taken into account in our model.

3.1 | Overall architecture

Figure 3 shows the overall architecture of our system. The virtual human controller receives the information about where to look at from the engagement detection component and controls the gaze behavior of the character. The dialogue of the character is based on AIML Pandorabots*. For speech recognition, we use Google Speech Recognition.

We developed the virtual character in Unity 3D game engine. The 3D model is created in Daz3D†. The character has the capability of speaking, gazing, displaying facial expressions, conversational gestures, and idle animations. Lip synch and the low-level gaze movement are based on third-party assets from the Unity Asset Store‡§. Gestures are recorded

*http://www.pandorabots.com/
†http://www.daz3d.com/
‡http://lipsync.rogodigital.com/
§http://tore-knabe.com/unity-asset-realistic-eye-movements

FIGURE 4 Virtual human Sara doing a waving hand gesture

TABLE 1 Features for engagement detection

Feature	Explanation
Distance	Whether the user is close to the virtual human
Velocity	Whether the user is staying still or moving
Body Rot.	Whether the user configures his or her body towards the virtual human
Head Rot. (Horizontal)	Whether the user turns his or her face towards the virtual human (horizon.)
Head Rot. (Vertical)	Whether the user turns his or her face towards the virtual human (vertical.)
Field of view	Whether the user is close to the center of field of view of the virtual human
Speaking (sound)	Whether the user is speaking based on sound localization
Speaking (mouth)	Whether the user is speaking based on mouth movement

with a Vicon Motion Capture system and applied to our character. Facial expression and visemes are exported as blend shapes from Daz3D. The synchronization between speech, gaze, facial expressions, and gestures is realized using the Behavior Mark-up Language (BML).[13] For this, we developed a BML Realizer for Unity[¶]. Figure 4 shows the virtual character performing a waving hand gesture.

3.2 | Engagement detection

Engagement of the users might depend on various features such as the distance of the users to the virtual character, their posture, or velocity. Previous work considered different combinations of features. In our work, we use a wide range of features and analyze which features have higher importance for engagement detection. We calculate engagement on the basis of a contribution of multiple features derived from a Kinect depth camera. Table 1 shows the list of features that we find most important based on the social sciences literature and previous work.

[¶]https://www.staff.science.uu.nl/ yumak001/UUVHC/index.html

FIGURE 5 Features and engagement score shown on the Kinect stream

Given $k \in [1, n]$, $i \in [0, 5]$, $f_k^i(t) \in [0, 1]$, $w_k \in [0, 1]$, and n being the number of features, engagement $E^i(t)$ is calculated as below:

$$E^i(t) = w_1 f_1^i(t) + w_2 f_2^i(t) + \ \ldots \ + w_n f_n^i(t).$$

$f_k^i(t)$ is the normalized value of feature k at time t for person i. w_k is the weight (importance) of the feature. Figure 5 shows the features calculated based on the Kinect stream. In the following sections, we describe in detail how each feature is calculated.

3.2.1 | Distance (proxemics)

Research in social sciences investigated how people manage distance during social interactions.[14] It is considered along four zones: intimate zone (0 to 0.15 m), close intimate zone (0.15 to 0.45 m), personal zone (0.45 m to 1.2 m), social zone (1.2 to 3.6 m), and public zone (more than 3.6 m). Depending on the relationship and the context, people dynamically adjust their distances. In the context of engagement, distance is a major signal for conversation initiation. Most of the previous work assumed that engagement is higher when the user is closer to the virtual human or robot. We also started with that assumption and also included the effect of other modalities. For example, a person standing far away might also be engaged if there are other signs of engagement (e.g., greeting from far away). However, our observations show that in most cases, engaged users have a tendency to come closer.

In our system, the closer the user is to the virtual character, the more engaged he or she is. Given $i \in [0, 5]$ as the user id, X_i, Z_i as the user's coordinates in the Kinect space, $D_i = \sqrt{X_i^2 + Z_i^2}$, $D_{max} = 4.5\ m$ and $D_{min} = 0.5\ m$, the feature is calculated as $f_{dis}^i = 1 - (D^i - D_{min})/(D_{max} - D_{min})$. Figure 6 shows D_{min}, D_{max} and distances for two users, where $D_1 > D_2$. Therefore, the user on the right who is closer to the virtual character is expected to have higher engagement.

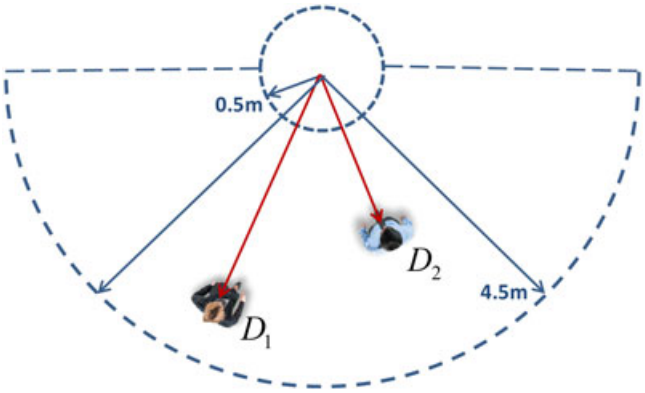

FIGURE 6 Distance to the virtual human

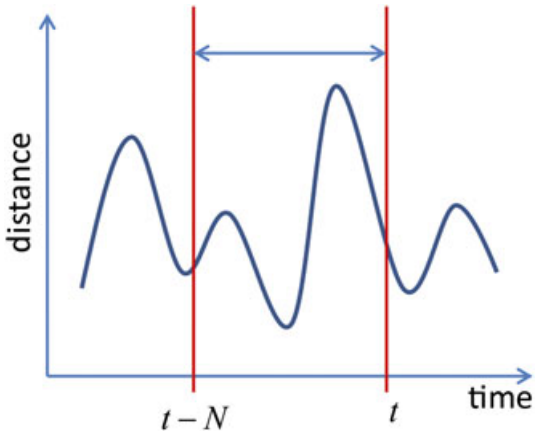

FIGURE 7 Velocity

3.2.2 | Velocity

Velocity is also an important measure in engagement. Even though the user is close to the virtual character for some frames, he or she might be a passer-by and not really trying to initiate a conversation. If the user is standing still, he or she is more likely to engage with the virtual character. We calculate the velocity over a running window of N frames. Figure 7 shows the window with N frames where the velocity is calculated for the frame at time t.

$$f^i_{vel} = 1 - \left(\frac{1}{N} \left(\sum_{j=t-N}^{j=N} = D^i_j - D^i_{j-1} \right) \times 30 \right) / 2.5.$$

We assume that maximum walking speed of a person is 2.5 m/s. Given N is the length of the window, D^i_t the distance in the current frame, and D^i_{t-N} the distance at the first frame in the window, f^i_{vel} is calculated as below, given the frame rate is 30 frames/s.

3.2.3 | Body rotation

If the user configures his or her body towards the virtual human, the user is considered more engaged. We calculate the posture direction of a user by finding the normal of the plane passing from the two shoulder joints and the spine joint. We also calculate the directional vector from the user's head towards the Kinect. The angle between these two vectors is called "body orientation deviation," represented as θ. The user is most engaged when θ is 0° and when f^i_{brot} takes its maximum value 1. As the user turns to the left/right, θ gets larger. We set the limitations of body rotation to ±45°. Therefore, the body orientation feature f^i_{brot} becomes 0 when θ is 45°. Figure 8 shows the angles between the posture vectors $\vec{P}_1, \vec{P}_2$ and the directional vector from the users to the Kinect. Because $\theta_1 < \theta_2$, the user on the left is considered more engaged.

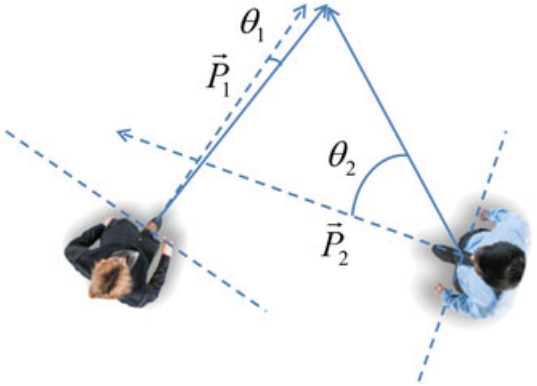

FIGURE 8 Body rotation

3.2.4 | Head rotation (horizontal)

Head orientation on the horizontal plane is calculated similar to the body orientation. We call the angle between the head rotation vector and the identity vector "head rotation deviation in the horizontal plane," which is represented as α. The feature takes its maximum value 1, when α is 0°. If the user turns to the left or right, the angle increases and engagement decreases. Setting the limits of head tracking to ±45°, the horizontal head orientation feature f^i_{hroth} becomes 0, when α is 45^o.

3.2.5 | Head rotation (vertical)

If the user is facing the virtual character, head rotation vertical feature takes its maximum value. The angle between the head vector up or down and the identity vector, in other words "head rotation deviation in the vertical plane" is defined as β, which takes its minimum value 0° when the user is facing the virtual character. It gets higher as the user moves his or her head up or down. Thus, the vertical head orientation feature f^i_{hrotv} is 1, when β is 0°, and it becomes 0, when β is 45°.

3.2.6 | Closeness to the center of field of view (FoV)

Apart from the orientation of the user, we look at whether the user is close to the center of FoV of Kinect. That is

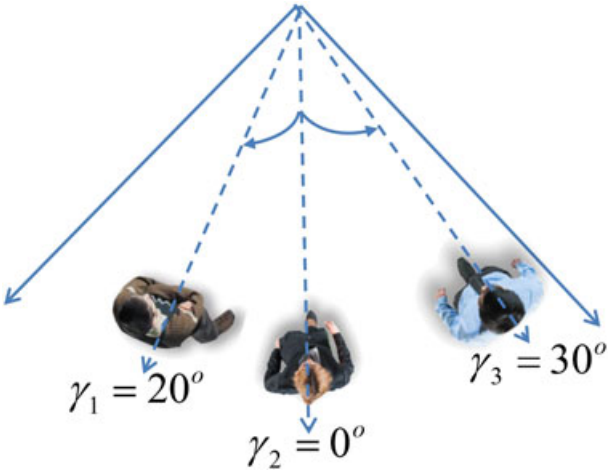

FIGURE 9 Closeness to the center of field of view

calculated based on the directional vector from the Kinect towards the user. We call the angle between the identity vector and the directional vector "field of view deviation," which is represented as γ. The more the user is at the center, the more engaged the user is considered. When the user is at the center, γ is equal to 0°. When the user is at the borders of Kinect's FoV (70°), it becomes 35°. Therefore, the feature f^i_{FoV} takes the value 1 when the user is at the center, and it becomes 0 when he or she is on the sides. Figure 9 shows the FoV deviation for three users, where $\gamma_3 > \gamma_1 > \gamma_2$. Therefore, the user in the center is considered to be the most engaged.

3.2.7 | Speaking (from sound localization)

On the basis of the sound angle beam returned from Kinect, we find the user that is estimated to be speaking based on the location of the user. To do this, we take into account the Kinect's visual FoV (±35°) and the Kinect's audio beam range (±50°). The angle ω between the directional vector of the audio source and the directional vector from the Kinect to the user is calculated. The user that has the smaller angle is chosen as the speaking user. In case of silence, the speaking value is set back to nonspeaking after 1 s. When a user is labeled as speaking, we calculate the feature value f^i_{speak} as 1, otherwise 0. Figure 10 shows the directional vector for the sound source with red and the directional vector to the users with blue. Because $\omega_1 < \omega_2$, the user on the left is considered as the speaking user and more engaged.

3.2.8 | Speaking (from mouth movement)

In addition to the sound angle, we look at whether the user is speaking based on the mouth movement. It takes three values $\{yes, maybe, no\}$. We map them to numeric values accordingly: $f^i_{mouth} = \{1.0, 0.5, 0.0\}$.

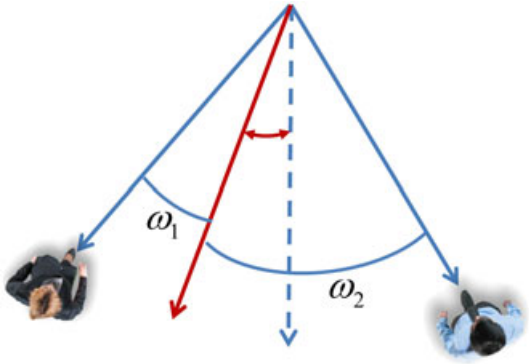

FIGURE 10 Speaking user

4 | EXPERIMENT

We conducted an experiment with the engagement model described above. We had two goals in mind: (a) finding out whether the model behaves well based on the data collected in the real environment and (b) finding an improved model based on weights learned from the data. Therefore, we have two models as described below:

- Base model: The feature weights are all equal.
- Improved model: The features weights are different and learned from data.

4.1 | Data collection and analysis

We installed the system at the entrance of our building next to the reception desk. The virtual character was rendered on a 46 in. screen to keep the virtual character close to life size. The set-up included a Kinect v2 for tracking the users and detecting the sound source. We also had a noise-cancellation microphone for speech recognition.

Table 2 shows the summary of collected data in each session. We collected data in five sessions. In total, there were 18 different subjects. Each session had three to seven users, where a few users appeared in multiple sessions. The amount of data collected was 31 min. The Kinect stream was recorded for further analysis and annotation of user behavior. In addition, we recorded the interaction with two video cameras from different angles. The final engagement score was calculated based on assigning equal weights to each of the features,

TABLE 2 Summary of collected data

Session	No. of users	Total data (minutes)	Annotated data (minutes)
1	7	7	6
2	3	8	2
3	3	7.5	2
4	3	4.5	1.5
5	5	4	2
Total	21	31	13.5

TABLE 3 Summary of coefficients and significance for the runs in the leave-one-out cross validation

			Features							
Test	**Training**		**Speaker**	**Body Rotation**	**Distance**	**FoV**	**Head Rot. H**	**Head Rot. V**	**Mouth Movement**	**Velocity**
1	2,3,4,5	Coeff.	0.487	**2.692**	4.449	−1.195	**0.746**	1.065	**−0.186**	**1.057**
		Signi.	0.196	**0.000**	0.102	0.313	**0.019**	0.107	**0.007**	**0.017**
2	1,3,4,5	Coeff.	0.150	**3.096**	**9.103**	**−1.517**	**0.929**	1.025	**−0.209**	**3.523**
		Signi.	0.621	**0.000**	**0.000**	**0.018**	**0.007**	0.096	**0.044**	**0.002**
3	1,2,4,5	Coeff.	0.192	**2.924**	**6.558**	−0.047	**1.251**	1.415	**−0.260**	**2.176**
		Signi.	0.558	**0.000**	**0.000**	0.952	**0.000**	0.084	**0.000**	**0.009**
4	1,2,3,5	Coeff	0.023	**2.527**	**6.116**	−0.308	**1.674**	1.038	−0.282	**5.228**
		Signif.	0.92	**0.000**	**0.004**	0.724	**0.000**	0.304	0.068	**0.001**
5	1,2,3,4	Coeff.	0.198	**2.078**	**5.377**	0.517	**0.912**	0.671	**−0.235**	**2.793**
		Signi.	0.207	**0.000**	**0.000**	0.382	**0.000**	0.070	**0.002**	**0.005**

Note. FoV = field of view.

which we call the "Base model." The engagement score and the values of the features are stored for each frame.

We annotated the data using ELAN Video Annotation Tool[‖]. For each user ID assigned by the Kinect, an annotation track is created. The moments when a user is engaged with the virtual character is labeled as *engaged*, while the rest are labeled as *not-engaged*. In total, we annotated 13.5 min of data.

Because our goal is to find the importance of the features, we run a regression analysis on the data. The data from the computational model and the annotated data is synchronized and sampled every 250 ms. The data is organized per session and per subject in order to take into account the correlations with-in subjects. Because our annotated data (dependent variable) was binary (engaged or not-engaged), we used logistic regression instead of linear regression. Therefore, we used the generalized estimating equations (GEEs) to analyze our data, which is a logistic regression model for correlated data. We selected the AR(1) autoregression as the correlation structure.

In order to find the correlations between engagement predictions and the annotated data, we calculate the predictions based on GEE coefficients. Although in linear regression, coefficients and predictions have a linear relationship, in logistic regression, coefficients have a linear relationship with the log odds of the prediction. Therefore, we calculate the predictions based on the following formula, where $B_0, \ldots, B_n$ are the coefficients and $X_0, \ldots, X_n$ are the independent variables (or features). P is the probability of the engagement predictions.

$$P = \frac{e^{B_0+B_1X_1+\ldots+B_nX_n}}{1+e^{B_0+B_1X_1+\ldots+B_nX_n}}.$$

[‖] https://tla.mpi.nl/tools/tla-tools/elan/

4.2 | Results and discussion

We run a leave-one-out-cross validation. At each run, one session was used as the test data and the remaining data was used for training. Table 3 shows the summary of coefficients and significance values for each feature. The values that are significant are marked as bold ($p < .05$).

The predictions from each run were used to find out whether the GEE model (coefficients-based model) is better in comparison to the base model (equal weights model). We found that both models correlates significantly with the ground-truth data, and the GEE-based model had a better correlation ($r = 0.437$, $p < .05$) in comparison to the base model ($r = 0.389$, $p < .05$)

This shows that assigning importance weights to the features gives a more accurate calculation for the final engagement score. In order to find the final set of coefficients, we run GEE on the whole data set including five sessions. We found that body rotation, distance, horizontal head rotation, mouth movement, and velocity have significant effects on the model ($B_2 = 2.604$, $p < .05$, $B_3 = 6.033$, $p < .05$, $B_5 = 1.115$, $p < .05$, $B_7 = -0.259$, $p < .05$), $B_8 = 3.148$, $p < .05$, respectively). Speaking, FoV and vertical head orientation features were found to be insignificant ($B_1 = 0.224$, $p > .05$, $B_4 = -0.312$, $p < .05$, $B_6 = 1.125$, $p > .05$). Table 4 shows the coefficients and significance of the features found after running GEE analysis on the whole data set.

Vertical head orientation has a positive coefficient, but it did not have a significant effect on the model. This might be due to the fact that the data did not contain too many up or down head movements. Initially, we added this parameter to consider the cases where users look down, that is, to look at their phone while they are staying idle. However, in our data, these cases were not observed frequently and users mostly had their face oriented towards the virtual human.

TABLE 4 Coefficients and significance values based on all data

Feature	Coefficient	Significance
Speaking	0.224	0.408
Body Rotation	**2.604**	**0.000**
Distance	**6.033**	**0.000**
FoV	−0.312	0.638
Head rotation horizontal	**1.115**	**0.001**
Head rotation vertical	1.125	0.076
Mouth movement	**−0.259**	**0.003**
Velocity	**3.148**	**0.001**

Note. FoV = field of view.

Speaking and FoV features had the least significance. The reason that the speaking feature is insignificant might be due to the Kinect sensor. Sound beam angle returned by the Kinect API is not only based on human speech but also based on other sources of sound in the environment. Regarding the FoV feature, when the users were interacting in groups, they were marked as engaged regardless of whether they were away from the center. Because the number of cases where the users were both on the sides and engaged were higher, that might express this effect.

In order to see the effect of our final model, we compared the predictions from the Base model with the GEE model. Both models correlated significantly with the annotated ground truth data. The GEE model has a higher correlation value ($r = 0.522$, $p < .05$). We found that the final model is an improved model with respect to the base model ($r = 0.389$, $p < .05$).

5 | CONCLUSION AND FUTURE WORK

In this paper, we described an engagement-driven gaze model for an interactive virtual character. We tested our model in a real environment and found that our initial model performs well when feature weights are all equal to each other. Our findings show that distance, velocity, body, horizontal head rotation, and mouth movement have significant effects on the model. Sound, vertical head orientation, and FoV parameters did not behave as we initially expected.

There are also limitations of our work. In order to improve the reliability of the annotations, multiple annotators can be employed by looking at the correlations among the annotators. Instead of using binary engagement labels, annotations can also be done over a range. It will also be interesting to apply and compare other machine learning models. Finally, further data collection and analysis can be done to capture various combinations of user behaviors. Although our results provide useful insights in terms of the importance of feature weights, a validation experiment should be run in order to see whether the new model is more socially adept. This paper is a first attempt to collect and analyze real-life data for engagement detection in a truly open space.

ACKNOWLEDGEMENTS

This work is supported by the Horizon 2020 RAGE—Realizing an Applied Gaming Eco-system project (Grant no. 644187) and Utrecht University Game Research Seed Money. We like to thank Dr. Ad Feelders for his feedback on the statistical analysis.

REFERENCES

1. Ruhland K, Peters CE, Andrist S, et al. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. Comput Graph Forum. 2015;34(6): 299–326.
2. Sidner CL, Lee C, Kidd CD, Lesh N, Rich C. Explorations in engagement for humans and robots. Artif Intell. 2005;166(1–2):140–164.
3. Michalowski MP, Sabanovic S, Simmons R. A spatial model of engagement for a social robot. 9th IEEE International Workshop on Advanced Motion Control, 2006, IEEE; 2006. p. 762–767.
4. Bohus D, Horvitz E. Learning to predict engagement with a spoken dialog system in open-world settings. Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '09, Association for Computational Linguistics; Stroudsburg, PA, USA; 2009. p. 244–252.
5. Foster ME, Gaschler A, Giuliani M. How can i help you? comparing engagement classification strategies for a robot bartender. Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI 2013); Sydney, Australia; 2013. p. 255–262.
6. Yumak Z, Magnenat-Thalmann N. Multimodal and multi-party social interactions. In: Magnenat-Thalmann N, Yuan J, Thalmann D, You B-J, editors. Context Aware Human-Robot and Human-Agent Interaction. Cham: Springer International Publishing, 2016. p. 275–298.
7. Kendon A. Spacing and orientation in co-present interaction. Proceedings of the Second International Conference on Development of Multimodal Interfaces: Active Listening and Synchrony, COST'09. Berlin, Heidelberg: Springer-Verlag, 2010. p. 1–15.
8. Peters C, Castellano G, de Freitas S. An exploration of user engagement in hci. Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots, AFFINE '09, ACM; New York, NY, USA; 2009. p. 9:1–9:3.
9. Xu Q, Li L, Wang G. Designing engagement-aware agents for multiparty conversations. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, ACM; New York, NY, USA; 2013. p. 2233–2242.
10. Zaraki A, Mazzei D, Giuliani M, Rossi DD. Designing and evaluating a social gaze-control system for a humanoid robot. IEEE Trans Human-Machine Syst. 2014;44(2):157–168.
11. Kokkinara E, Oyekoya O, Steed A. Modelling selective visual attention for autonomous virtual characters. Comput Animat Virtual Worlds. 2011;22(4):361–369.
12. Grillon H, Thalmann D. Simulating gaze attention behaviors for crowds. Comput Animat Virtual Worlds. 2009;20(23): 111–119.

13. Kopp S, Krenn B, Marsella S, et al. Towards a common framework for multimodal generation: The behavior markup language. Proceedings of the 6th International Conference on Intelligent Virtual Agents, IVA'06. Berlin, Heidelberg: Springer-Verlag, 2006. p. 205–217.
14. Hall ET. The hidden dimension, USA: Anchor Books, 1966.

Dr. Zerrin Yumak is an assistant professor in Computer Science at Utrecht University. Her research is on socially intelligent virtual humans. She worked on emotions and memory modeling, multiparty interac tions, facial animation and gaze generation for virtual characters, and social robots. Zerrin received a PhD degree in Computer Science from MIRALab, University of Geneva. Following that, she did a postdoc at the HCI Group in EPFL, Switzerland, and worked as a research fellow at the Institute for Media Innovation, Nanyang Technological University in Singapore. She is a program committee member of the CASA and CGI conferences and editorial board member of the Visual Computer journal. She has been involved in several EU and national funded projects. Currently, she is leading Task 3.2: Embodiment and Physical Interaction in the RAGE (Realising an Applied Gaming Eco-system) EU project.

Bram van den Brink is a student in the Master Program Game and Media Technology at Utrecht University. He also works as a part-time game developer at Vogelsap. His interests are modeling and animation, physics engines, shaders, and game mechanics.

Dr. Arjan Egges is an associate professor at Utrecht University, where he founded the Virtual Human Technology Lab in 2013. His research focuses on different aspects of character simulation and animation. He has investigated automatic motion concatenation systems such as motion graphs or snap-together motions. He has worked on real-time crying simulation and rendering, example-based motion synthesis, and emotion simulation. Furthermore, he is performing several studies on the perception of motion. Arjan teaches several courses related to game programming and computer animation. Arjan is the cofounder of the annual Motion in Games conference. He is a member of many program committees, including CASA, CGI, SCA, IVA, PG, GRAPP, and more. Over the last years, Arjan was involved in obtaining several research grants, among them are COMMANDS project (Agentschap NL), COMMIT (virtual worlds for well-being) P4 national project, and TARDIS (STREP EU project). Arjan is currently CTO at Fans4Music.

How to cite this article: Yumak Z, van den Brink B, Egges A. Autonomous social gaze model for an interactive virtual character in real-life settings. *Comput Anim Virtual Worlds*. 2017;28:e1757. https://doi.org/10.1002/cav.1757