Increasing Engagement with Virtual Agents Using Automatic Camera Motion

Lazlo Ring, Dina Utami, Stefan Olafsson, Timothy Bickmore

College of Computer and Information Science, Northeastern University, Boston, MA
{lring, dinau, stefanolafs, bickmore}@ccs.neu.edu

Abstract. We describe a series of algorithms which automatically control camera position in a virtual environment while a user is engaged in a simulated face-to-face dialog with a single virtual agent. The common objective of the algorithms is to increase user engagement with the interaction. In our work, we describe three different automated camera control systems that: (1) control the camera's position based on topic changes in dialog; (2) use sentiment analysis to control the camera-to-agent distance; and (3) adjust the camera's depth-of-field based on "important" segments of the dialog. Evaluation studies of each method are described. We find that changing camera position based on topic shifts results in significant increases in a self-reported measure of engagement, while the other methods seem to actually decrease user engagement. Interpretations and ramifications of the results are discussed.

Keywords: relational agent, cinematography, natural language understanding.

1 Introduction

As we develop agent-based interfaces for education, healthcare, and entertainment, maintaining user engagement represents a growing area of concern [1]. While there has been some effort to increase engagement through the manipulation of an agent's dialog, little work has been done to explore how agents can keep a user engaged without authoring major extensions to the dialog content of the system. Many agent applications require designing voluntary-use systems for long-term interaction, or maintaining the sustained attention of the user, such as automatic health behavior change systems or life-long learning companions [2]. Maintaining engagement with the user can be crucial for these systems to succeed, since engagement is often a prerequisite for other system objectives: if a user stops interacting with a system, then it cannot have any further impact.

While some researchers have explored the use of superficial variability in linguistic choice to increase user engagement [3], or the use of agent backstories [4], storytelling [5], or appropriate agent listening (backchannel) behaviors [6], these changes require significant work on the part of developers and content writers. We believe that the field of Cinematography offers insights that can be automatically integrated into a virtual agent application to increase user engagement.

By drawing from cinematographic principles, an automated camera control system that automatically adjusts the user's view of the agent can be developed, based on linguistic analysis, in order to increase engagement. While prior research has explored the creation of such systems in the past [7], they have primarily focused on multi-agent interaction rather than creating engaging one-on-one conversations between a user and a single agent.

In this paper, we explore the design and creation of such a system, by developing and evaluating an automated camera system focused around maintaining user engagement in one-on-one conversations, independent of dialog content. Based upon cinematographic theory, the camera system uses natural language processing to automatically control the user's camera during a virtual conversation. We also evaluate various camera manipulation techniques through a series of sub-studies that explore their potential effects on user engagement.

2 Related Work

In this section we review prior work on automated camera control in 3D virtual environments, in particular for a single stationary actor in the scene.

De Melo and Paiva presented a model for automatic manipulation of the light and screen expression channels. The model integrates the OCC emotion model for emotion [8], expressively controls lights and shadows, and looks to visual arts techniques for layering and filtering to manipulate a virtual agent. Their main manipulation of the camera was using proxemics: either dollying or zooming in to increase drama [9].

Canini, et al, developed a model that could estimate camera-subject distance in film, using image processing and machine learning algorithms. The system performed with over 80% accuracy and their results revealed that the director can impact the perceived affective response of the viewer, caused by alternating between close-up, medium, and long shots, with close-ups having the highest level of arousal. There was no connection between shot type and emotional valence [10].

Rui, et al, developed an automated videography system for lectures being broadcast to remote audiences, by coming up with rules for best practices to make videos visually engaging. The rules for camera positions included (1) camera placement and angle; (2) and cameras should be close to eye level. The rules for shot transitions included (1) reasonably frequent shot changes, (2) defining a minimum shot duration, (3) shot transitions should be motivated, and (4) transition when the speaker finishes a concept or thought [11].

Calahan describes various methods for lighting and other effects in computer graphics that have been found to enhance visual storytelling. Perspective and depth of field can be manipulated by changing the focal point or blurring particular planes in the scene, e.g. the background, which will emphasize the subject in the foreground, creating a greater sense of intimacy than one without blur [12].

3 Exploring Automatic Camera Control

Based on prior work, we designed a series of increasingly sophisticated methods for automatically controlling a camera while a user is engaged in a simulated face-to-face dialog with a single virtual agent. Our overarching goal was to develop methods that only required the text of the utterances that the agent will speak; we derived camera control parameters from linguistic analyses of the user-agent dialog script, the discourse history, and a description of the current virtual scene. We evaluated each of our methods in 3-treatment, between-subjects experiments where user engagement was the primary means measure of impact. Before describing each of the camera control methods and individual evaluation studies, we first describe the experimental methods that are common to all of them.

3.1 Common Study Methods

Each evaluation study was a between-subjects experiment with three treatments: the automated camera control algorithm being evaluated (AUTO); an equivalent agent interaction with no camera motion (STATIC); and a condition in which the same camera controls used in a representative run through the AUTO interaction were used, but deployed at random points in the dialog script (RANDOM).

In each study, participants interacted with a female virtual agent that spoke using synthetic speech and synchronized nonverbal conversational behavior automatically generated by BEAT [13] (Figure 1). User inputs to the dialog were made via multiple choice inputs updated at each turn of the dialog.

The dialog script used for all studies comprised 6 turns of social chat, followed by 15 turns of "task talk", during which the agent attempted to persuade the user to get more exercise, followed by 18 turns of social chat, in which the user was given the ability to end the conversation at every turns.

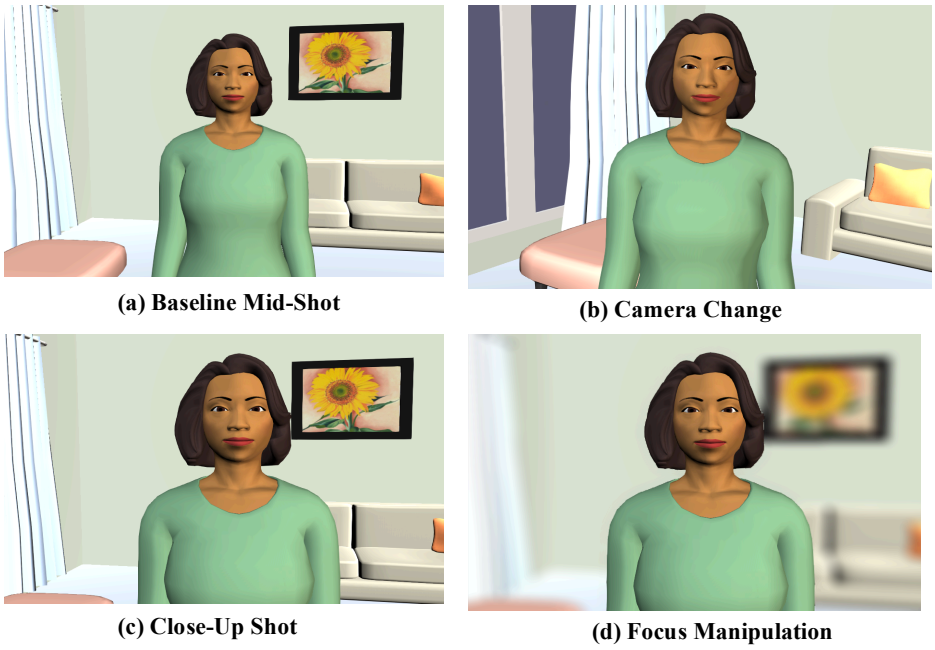
The evaluation studies were all conducted on Amazon's Mechanical Turk (AMT). All participants were required to have a 75% or higher approval rating on AMT, with the only additional requirement being that they had to use either Firefox or Chrome as their web browser. Participants who accepted the HIT from Mechanical Turk were asked to completed a socio-demographic and exercise attitude questionnaire, and then engage in a 5-10 minute conversation with the agent.

Upon completing the interaction, participants were presented with a series of questionnaires assessing their level of engagement, motivation to exercise, and overall impression of the system.

3.2 Common Study Measures

Engagement was our primary outcome of interest, and we measured it using both self-report and behavioral methods.

We developed a 26-item composite scale self-report measure of engagement (Table 1). Items were included from several prior studies and measures [14, 15], forming a pool of 31 items. Following our first study with 284 participants (Section 4), we conducted a factor analysis and found that one factor explained 36.4% of the variance, and we retained this one factor as our single measure of engagement, dropping 5



(a) Baseline Mid-Shot **(b) Camera Change**

(c) Close-Up Shot **(d) Focus Manipulation**

Fig. 1. Camera Manipulations Evaluated

items. The final measure had adequate internal consistency, with Cronbach's alpha ranging from .961 to .969 across the three studies.

We also measured engagement behaviorally, by recording the total number of dialog turns which users conducted with the agent, including the 18 optional turns of social chat at the end of each interaction.

Finally, we wanted to determine if engagement could play a mediating role in the agent's ability to change user attitudes towards exercise, as a task outcome measure. We measured the user's exercise stage of change using a validated self-report instrument [16], along with two additional questions about motivation and confidence to exercise, i.e. "How motivated are you to exercise more than you are currently" and "How confident are you that you could exercise more if you wanted to," using single item measures, all administered at the beginning and end of each interaction.

Participants were also given the chance to write additional comments about the agent into a text box following their interaction.

4 Approach 1: Changing Camera on Topic Shift

Our initial approach to automating camera motion followed observations of common practice in newscasts and commercial videotaped lectures (see, for example: www.thegreatcourses.com), in which an occasional camera change is used to increase visual variety when a single character is speaking on camera for an extended length of time.

Table 1. Engagement Questionnaire

Question		Anchor 1	Anchor 7
The character's behavior was natural. I felt like I was talking face-to-face with a person. The character's behavior was comfortable. The character's motion was pleasant. I could easily understand the character. I felt comfortable interacting with the character. The character was engaging. The character was charismatic. The character was warm.	I would like to interact with the character again. I had fun interacting with the character. I enjoyed interacting with the character. I found the character was entertaining. I liked interacting with the character. I was energized by my interaction with the character. I was alert during my conversation with the character. I felt the conversation was too short	Disagree completely	Agree completely
The character's behavior was repetitive. The character was weird. The character was boring.	I felt awkward talking to the character. I disliked interacting with the character.	Agree completely	Disagree completely
How friendly was the character?		Very unfriendly	Very friendly
How trustworthy was the character?		Very untrustworthy	Very trustworthy
How much do you like the character?		Not at all	Very much
How much do you feel that the character cares about you?		Not at all	Very much

We explored automatically changing camera position at topic boundaries in the agent's dialog, as prior studies have determined that a speaker's posture shifts are significantly more likely to occur during these transitions [17]. This is also supported by researchers in sociolinguistics who observed that changes in the spatial relationship between two speakers tended to occur at topic boundaries ("situational shifts") [18]. Topic boundary detection was performed automatically using the built-in mechanism in BEAT [13] that relies primarily on the identification of discourse markers [19] in the agent's script (e.g., "well", "anyway", "so", etc.).

This camera controller was implemented by using two cameras, 75 degrees apart relative to the agent in the virtual environment, and alternating between them when a camera change was indicated. Immediately following a camera change, the agent would turn to face the current camera.

4.1 Camera Change Evaluation Results

We had 284 participants, 58.1% male, between the ages of 18-69 (mean 34.8) participate in this study, with 102 randomized to the automatic camera condition (AUTO) condition, 88 to the random camera condition (RANDOM), and 94 to the static camera condition (STATIC).

A one-way ANOVA demonstrated that there were significant differences among study conditions on self-reported engagement, $F(2,281)=4.12$, $p<.05$, $D=.18$. Bonferroni post-hoc tests at the .05 significance level demonstrated that participants rated the AUTOMATIC condition as significantly more engaging than the other two conditions, and that there were no significant differences between the STATIC and RANDOM condition on engagement (Table 2).

Turn count data indicated that many participants completed either the minimum or the maximum number of turns possible, yielding a bimodal distribution. Non-parametric Kruskall-Wallis tests indicated no significant differences on turn count across the three study conditions, $p=.41$. There were also no significant differences in exercise stage, motivation, or confidence between the three study conditions. However, non-parametric bivariate correlations (Spearman's rho) indicated a significant positive correlation between the self-report measure of engagement and turn count, $rho=.213$, $p<.001$, increases in motivation to exercise, $rho=.195$, $p=.001$, and increases in exercise confidence, $rho=.163$, $p=.006$, indicating that there was some effect of engagement on these measures.

One of the items from our composite measure that we feel best captures the sense of naturalness of the interaction is "I felt like I was talking face-to-face with a person." Non-parametric tests on this item alone also demonstrated significant differences across conditions, Kruskal-Wallis, $p=.002$, with STATIC=2.89, RANDOM=3.38, and AUTO=3.74.

Table 2. Outcomes for Camera Change Study (mean (sd))

Measure	**STATIC** (N=94)	**RANDOM** (N=88)	**AUTO** (N=102)	**p**
Self-report Engagement	4.18 (1.25)	4.18 (1.14)	4.53 (1.08)	0.03*
Dialog Turns	31.12 (6.5)	30.48 (6.4)	31.45 (6.8)	0.41
Exercise Stage Change	0.06 (0.38)	-0.02 (0.37)	-0.01 (0.4)	0.13
Exercise Motivation Change	0.22 (1.53)	-0.10 (0.92)	-0.15 (1.15)	0.06
Exercise Confidence Change	0.03 (1.17)	-0.06 (0.75)	-0.16 (1.02)	0.13

4.2 Camera Change Evaluation Discussion

We demonstrated that changing the camera at topic changes led to significant increases in self-reported engagement, compared to a single static camera or a camera changed at random times. There was some evidence that this positively impacted task outcome measures (exercise motivation and confidence).

5 Approach 2: Adding Sentiment-based Camera Distance

Given our initial positive results, we attempted to increase the sophistication of our automated camera controller by investigating a mechanism for automatically adjusting the camera distance in addition to location. One of the other fundamental dimensions of camera control is the camera's distance to the agent, changing from "wide shots" to "extreme closeups" [20]. Camera distance has been investigated as a mechanism for partially indicating the "conversational frame" [21] in use by a virtual agent

(e.g., shifting between "task talk", "social talk", and "empathy talk" [22]). However, changes in conversational frame are infrequent, and automatic identification of frame or genre can be error prone.

Inspired by Canini's work on automatic affect sensing in cinematography [10] we investigated the creation of a camera controller that bases the camera's distance according to the emotional intensity of each agent-utterance, with the view that more emotionally intense utterances would be better received by the user if accompanied by close-up shots of the agent.

To generate sentiment ratings for the agent's dialog, we used the Stanford CoreNLP Toolkit [23] to label each agent utterance with one of five sentiment scores (Very Negative, Negative, Neutral, Positive, Very Positive), along with a probability rating. Our sentiment-based camera controller used a close up shot for utterances tagged as Very Negative or Very Positive, and a mid-shot used for all other utterances. We then conducted an evaluation study combining topic-based camera change controller (as evaluated in Study 1) with the sentiment-based camera distance controller.

5.1 Sentiment-based Camera Distance Results

We had 149 individuals, 56.4% male, between the ages of 19-65 (mean 36.4) participate in Study 2. Of the 149 participants, 51 were randomized to the automatic camera condition (AUTO) condition, 44 to the random camera condition (RANDOM), and 54 to the static camera condition (STATIC).

In the second experiment, we found no significant differences on self-reported engagement by study condition, $F(2,146)=1.31$, $p=.27$ (STATIC: Mean = 4.61, SD = 0.88, RANDOM: Mean = 4.32, SD = 1.01, AUTO: Mean = 4.19, SD = 0.91) . Trends in the data suggested that camera zoom had a negative impact on turn count, in which participants favored the static condition over the random and automatic camera conditions, Kurskal-Wallis, $p=.056$, with Means of STATIC (Mean = 34.3, SD = 7.25) > RANDOM (Mean = 32. 6, SD = 6.59) > AUTO (Mean = 31.9, SD = 6.69). Non-parametric correlation between engagement and turn count remained significant ($rho=.336$, $p<.001$.), however there were no significant result found for attitude change.

5.2 Sentiment-based Camera Distance Discussion

In this study we explored the use of a sentiment-based camera controller to improve user engagement. There were no significant effects of our automated camera control system on user engagement, although the trends suggested that the changes in camera motion may have actually lead to a more negative user experience. We found that the number of camera changes that occurred during the interaction was drastically higher than those found in study 1, with nearly 70% of all dialog turns containing at least one change in camera motion. This frequency of change may have conflicted with some of the best practice rules found in cinematography, namely that an overabundance of shot type changes can give the illusion that the director is "bored" [20].

6 Approach 3: Automating Camera Focus

Based on the qualitative feedback received during Study 2, we designed a subtler and less frequent camera change to signal emotional intensity. Rather than the jarring change in camera distance, we signaled high emotional intensity by manipulating depth of field to heighten the agent's contrast with her background (Figure 1d).

To reduce the frequency of these changes, we incorporated information about what parts of each topic or dialog segment represented the most "important" information. Relative importance of a given utterance is a function of the broader goals of a dialog (e.g., it could be the point of resolution in a narrative, the punch line to a joke, or the key message in a lecture). To simulate this, we manually added camera tags to the start, peak and end of each discourse segment. These tags allowed the camera system to automatically adjust the camera's distance and focus on the agent based on the various tags within each section. The closest distance and greatest focus occurred at the utterance tagged as the most important, with gradual transitions occurring into and out of this peak for at least two utterances on each side.

6.1 Camera Focus Results

We had 99 individuals, 51% male, aged 19-74 (mean 38) participate in the evaluation of the Camera Focus controller, with 37 randomized to the automatic camera condition (AUTO) condition, 34 to the random camera condition (RANDOM), and 28 to the static camera condition (STATIC).

As with Study 2, no significant differences were found between study conditions on engagement, $F(2,96)=.364$, $p=.696$ (STATIC: Mean = 4.38, SD = 1.25, RANDOM: Mean = 4.12, SD = 1.2, AUTO: Mean = 4.28, SD = 1.24). Similarly, there was also a trend suggesting that the changes in camera motion had a negative impact on turn count, Kruskal-Wallis, $p=0.11$, with Means of STATIC (Mean = 32.18, SD = 6.77) > RANDOM (Mean = 30.06, SD = 6.47) > AUTO (Mean = 28.97, SD = 5.6). A non-parametric test for turn count differences between STATIC and AUTO found the same trend (Mann-Whitney $U=363$, $p<.05$). The non-parametric correlations between self-reported engagement and other measures remained significant (for turn count: $rho=0.20$, $p<.05$; for change in exercise motivation, $rho=0.314$, $p=.002$; and for change in exercise confidence, $rho=0.239$, $p<.05$.)

6.2 Camera Focus Discussion

Building upon our results from the previous two studies, we developed an automated camera controller that adjusted the cameras distance and level of focus in relation to the importance of dialog utterances. As with study 2, there was no significant correlation between automated camera control and user engagement, with results trending against the automated system. A qualitative analysis of user feedback suggested that even though camera motion was less frequent than that of study 2 (42% of turns vs 70% of turns), the changes in camera proxemics was un-enjoyable.

7 Overall Discussion and Conclusion

In this paper we explored the creation of an automated camera system for conversational agent based systems. We explored the potential impact of three different automated camera systems, which adjusted camera position, proximity and focus, on user engagement and motivation. Our results demonstrated that automated camera motion, especially in relation to topic shifts, can have a positive impact on user engagement, while changes in proxemics and focus control trended towards having a negative effect. Additionally, we found a significant correlation between engagement, turn count, and the agent's persuasiveness.

This suggests the potential of an automated camera system for increasing user engagement and improving the effectiveness of agent-based systems. However, our studies demonstrated that randomly manipulating camera motion can have a negative impact on these metrics, and that any change in camera motion needs to be thoroughly studied before integration. Additionally, a large range of factors seem to contribute to the user's enjoyment of such systems, including motion speed, frequency of motion, and when the motion occurs in relation to the agent's utterances.

The results may be simply due to the fact that while a small amount of camera motion is important for users to maintain engagement with a single virtual agent, additional camera movement (of any kind) beyond a threshold becomes a distraction (Table 3).

Table 3. Comparison of Three Automatic Controllers (mean (sd))

	Camera Change	**Sentiment/Distance**	**Importance/Focus**
% Turns with any Camera Change	11%	70%	42%
Self-report Engagement	4.53 (1.08)	4.19 (0.91)	4.28 (1.24)
Dialog Turns	31.45 (6.8)	31.9 (6.69)	28.97 (5.6)

8 Limitations & Future Work

Although we explored three different types of camera manipulation within our study—position, distance, and focus—we did not explore sub-factors for each manipulation. As shown in the differences between study 2 and 3, minor changes in camera motion can greatly impact a participant's opinion of the system, suggesting that we should more thoroughly explore sub-factors such as the motion speed and overall frequency of camera motion. In future work, we plan to explore each of these sub-factors in a controlled environment and see how these effects persist in longitudinal interactions.

References

1. Lehmann, J., Lalmas, M., Yom-Tov, E., Dupret, G.: Models of User Engagement. User Modeling, Adaptation, and Personalization, vol. 7379, pp. 164-175. Springer (2012)
2. http://projects.ict.usc.edu/companion/
3. Bickmore, T., Schulman, D., Yin, L.: Maintaining Engagement in Long-term Interventions with Relational Agents. International J. of Applied AI, 24, 648-666 (2010)
4. Bickmore, T., Schulman, D., Yin, L.: Engagement vs. Deceit: Virtual Humans with Human Autobiographies. In: Intelligent Virtual Agents. (2009)
5. Battaglino, C., Bickmore, T.: Increasing Engagement with Conversational Agents Using Co-Constructed Storytelling. Int8 workshop, Santa Cruz, CA (2015)
6. Smith, J.: GrandChair: Conversational Collection of Grandparent's Stories. MIT (2000)
7. Christie, M., Machap, R., Normand, J., Olivier, P., Pickering, J.: Virtual Camera Planning: A Survey. 5th International Symposium on Smart Graphics, (2005)
8. Ortony, A., Clore, G., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, MA (1988)
9. De Melo, C., Paiva, A.: Expression of emotions in virtual humans using lights, shadows, composition and filters. Affective Computing and Intelligent Interaction pp. 546-557. Springer, Berlin Heidelberg (2007)
10. Canini, L., Benini, S., Leonardi, R.: Affective analysis on patterns of shot types in movies. 7th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 253-258 (2011)
11. Rui, Y., Gupta, A., Grudin, J.: Videography for telepresentations. CHI'03, pp. 457-464 (2003)
12. Calahan, S.: Storytelling through lighting: a computer graphics perspective. (1996)
13. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: SIGGRAPH '01, pp. 477-486. (2001)
14. Nowak, K., Biocca, F.: The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. Presence 12, 481-494 (2003)
15. DeVault, D., Mell, J., Gratch, J.: Toward Natural Turn-Taking in a Virtual Human Negotiation Agent. AAAI 2015 Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction,
16. Marcus, B., Simkin, L.: The stages of exercise behavior. J Sports Med Phys Fitness 33, 83-88 (1993)
17. Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., Rich, C.: Non-Verbal Cues for Discourse Structure. In: Association for Computational Linguistics, pp. 106-115. (Year)
18. Blom, J., Gumperz, J.: Social Meaning in Linguistic Structures: Code Switching in Northern Norway. In: Gumperz , J., Hymes, D. (eds.) Directions in Sociolinguistics. Holt, Rinehart, and Winston, New York (1972)
19. Schiffrin, D.: Discourse markers. Cambridge University Press, Cambridge (1987)
20. Arijan, D.: Grammar of the Film Language. Silman-James, Los Angeles (1976)
21. Tannen, D. (ed.): Framing in Discourse. Oxford University Press, New York (1993)
22. Bickmore, T., Picard, R.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer Human Interaction 12, 293-327 (2005)
23. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 55-60 (2014)