第 26 回日本バーチャルリアリティ学会大会論文集（2021 年 9 月）

A Model for Virtual Guiding Agent based on People's Visual Attention Estimation for Actively Open Interactions

陳品蓉 [1)]，三武裕玄 [1)]，長谷川晶一 [2)]
Pinjung CHEN, Hironori MITAKE, and Shoichi HASEGAWA

1) Tokyo Institute of Technology （〒 152-8550 2-12-1 Ookayama, Meguro City, chen.p.ad@m.titech.ac.jp）
2) Tokyo Institute of Technology （〒 152-8550 2-12-1 Ookayama, Meguro City, mitake@haselab.net）
3) Tokyo Institute of Technology （〒 152-8550 2-12-1 Ookayama, Meguro City, hase@haselab.net ）

概要: Interactive agents are gradually used in guidance like travel information guides. However, agents are usually passive, so that people are forced to suspend their current tasks and request agents with explicit demand. It is necessary to make agent more actively open the interaction naturally but without being bothering. To achieve the goal, it is vital to distinguish whether people need explanation and the target to talk about. We use eye-tracking device to estimate attention amount and create a model includes perception of players' visual attention amount toward the surroundings to know players' things of interest. Then, we determine reacting time based on observation from Wizard-of-Oz experiment.

キーワード： Visual attention, Gaze awareness, Guide, HAI

1. Introduction

Nowadays, virtual agents appear in our life more and more frequently along with the growing use of virtual reality entertainment or technology concierge tools. To have mutual interaction, agents can recognize people' needs or intentions by the input of their behavior and respond. However, agents are more likely to be passive in human-agent interaction, so people have to take the initiative to ask for help. Also, people have to suspend their current work to search for the agent. Therefore, it is necessary for agents to be proactive but not being disturbing at the same time.

To let agents take the initiative, they need to observe one's state carefully. To do so, humans' eye gaze is an implicit expression of a person. There are some benefits for agents to perceive human needs by gaze. Firstly, a proper opening to interact with people can be found by gaze signals. Secondly, listeners' engagement can be estimated in the conversation or interaction. Furthermore, agents can be more interactive. Therefore, recognition of gaze allows agents to be more aware of people and assist them.

The objective of our research is to allow the agent to recognize human's need based on their visual attention to the surroundings and assist them. We aim to mimic the reality by applying it on automatic agent guiding after perceiving priority of interest and determine what and how deeply to introduce to the player.

We propose a model to estimate the human's visual attention and distinguish interest of things. Furthermore, by collecting information in human-human interaction, the agent is able to determine the timing of reaction. Finally, the model is developed in VR environment to be experienced by human.

2. Related Works

Inference of visual attention is often applied in agent cooperation to assist people. By using POMDP[1], shared autonomy can be realized so that the assistance of robot can be done using indirect input. However, the human has strong task oriented inference of gaze using indirect input which the agent doesn't have to be hesitate to find a suitable timing for interaction.

Peters et al.[2] proposed a model using the perception of attention to be applied in a two people dialog. However, their model only contains people's engagement toward the speaking counterparts. The interaction with the surroundings is not included while perception of people's attention for the environment is not obtained.

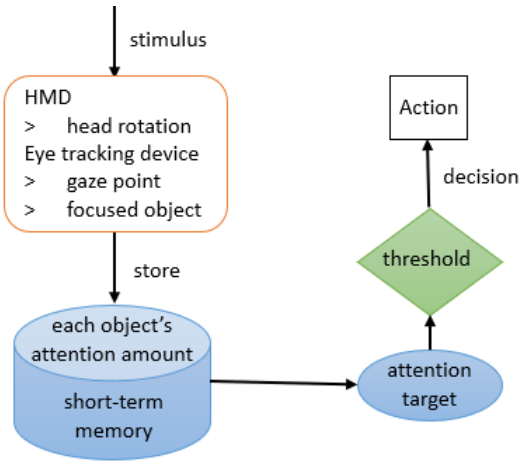

図 1: System Overview

Kevin et al.[3] developed an interaction method using rule-based AI for perception of gaze for both surroundings and agent. Nevertheless, it doesn't consider much about when the agent should react. In addition, the awareness of different attention toward the surrounding objects are not included.

3. Proposed Method

3.1 Estimation of Visual Attention Amount

We apply the method relating to Peter et al.'s module[2] for calculating attention amount. Besides storing the perception of attention amount toward the participant involved, our proposed method also allow agent to observe player's attention for the surroundings. Therefore, things of interest is known by comparison of attention amount.

3.1.1 Perception of attention

Sensed information is stored in short-term memory(STM) with 200ms interval to form perception of attention:

- head-eye angle(a_{he}): angle between head direction and gaze direction for the visual field in horizontal
- focused object(f_{obj}): current object player gazing
- index in short-term memory (i): order in perception of memory

In each memory slot, each object's angle information is calculated as the equation in 1 and stored in slot S_{visual}. g is the gain in sigmoid function.

$$S_{visual}(obj) = \begin{cases} 1 - (\frac{2}{1+e^{-ga_{he}}} - 1), & if\ f_{obj} = obj \\ 0, & if\ f_{obj} != obj \end{cases} \quad (1)$$

After getting each object's attention amount in memory slots, attention amount($A_{attention}$) is accumulated in memory as shown in equation 3. $t_{w_{t-i}}$ refers to the memory decay along with the time and we multiply the time sequence with linear coefficient. t represents the interval slots number containing in STM.

Then, we get the attention target($T_{attention}$) from maximum attention amount within the objects so that the agent has the preliminary distinguish of attention target as shown in equation 2.

$$A_{attention}(obj) = \sum^{t} 0.2 t_{w_{t-i}} S_{visual}(obj) \quad (2)$$

$$T_{attention} = \arg\max_{obj} \ (A_{attention}(obj)) \quad (3)$$

3.2 Interactivity threshold

We set a interactivity threshold that for the agent to decide the reacting time. It is shown in equation 4. The interactivity threshold(I) involves a base threshold(B) and the total gaze count(C_{gaze}) of the player stored in the agent's memory. W_c is the weight for C_{gaze}. Another important factor that affect the threshold is the direct gaze(D_{gaze}) from player to agent. It is an eye cue for the intention of needing help. Therefore, the threshold decrease in this circumstance. We set D_{gaze} as 0.4.

$$I = \begin{cases} W_c C_{gaze} + B, if\ there\ is\ no\ direct\ gaze \\ (W_c C_{gaze} + B) D_{gaze}, if\ there\ is\ direct\ gaze \end{cases} \quad (4)$$

3.3 Interaction Model

Figure 2 is the state machine diagram of how the agent's state transits within the interaction with the player. The model is driven by attention amount and interest threshold. The model contains start, observation, approach, introduction, and get back states. The trigger conditions of states are indicated in the figure.

3.4 Obtainment of Parameter

To obtain parameter of g, t, W_c, and B, we collect data from Wizard-of-Oz(WoZ) approach to measure natural human visual attention awareness interaction in reality. The guidance of introduction is reproduced by two people acting as a visitor and a guide(wizard). The wizard observes player's gaze and explains the target which player may be interested in. The data of attention amount is collected in the whole process. People's reaction and voice is recorded down by video. After WoZ experiment, we optimize parameters by applying covariance matrix adaptation evolution strategy(CMA-ES) to identify parameters. In the objective function, we calculate the relative error of reality-simulated time difference for the final score of relative error with penalty. It is the difference between reality annotated reaction time and

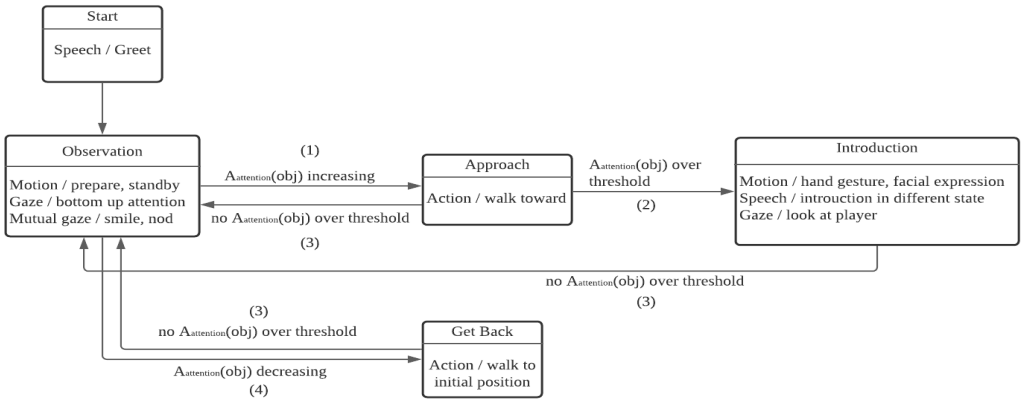

図 2: Interaction State Machine Diagram

図 3: Reproduction of Guiding Introduction using VGentWorld[4]

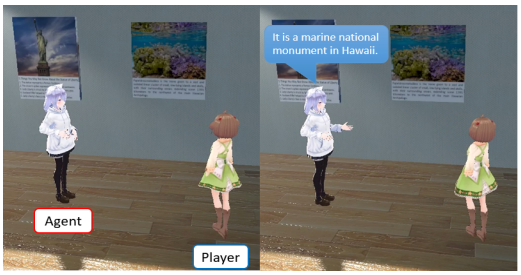

図 4: Introduction State

simulated reaction time. The penalty is the false judgement of reaction of the simulated reaction. Finally, the optimized values are utilized in our interaction model. From the optimization we can get the parameter value for: Base thresh(B):68.56, Count weight(W_c): 0.79, Angle gain(g):0.043, STM storage period(t):26 seconds.

4. Implementation

The proposed model is implemented in a conversational agent in VR. We use HTC VIVE Pro EYE for HMD and eye-tracking device. For action generation, VGentEditor[5] enables us to develop action in Unity.

5. Evaluation

We evaluated our model by user test comparing with two other models:

PROPOSAL Proposed Model

SIMPLE Simple gaze trigger from continuous gaze

WOZ Controlled by human

Participants experience the interactive agents utilizing different models in our created VR situation. In the experiments, they are asked to look around pictures in the exhibition and the agents will introduce for them. The impression of the models will be judged in three aspects which are assistance, bother, and naturalness. We ask the participants 8 questions using 7-point Likert scale and 2 questions for user feedback.

Q1. I think I get assistance from guide.
Q2. The guide can react to my interest precisely.
Q3. I feel the guide response in time.
Q4. I feel the guide response too late.
Q5. I feel the guide response too quick.
Q6. I feel bothered by guide.
Q7. I feel tired after the visit.
Q8. I think the guide behaves naturally.
Q9. In what aspect do you feel unnatural of agent.
Q10. What do you think of agent's performance?

There are 10 participants joining the experiment who are able to understand daily conversation in English. The participants are aged between 10-40(5 female, 5 male).

Figure 5 shows the result of impression by participants. By t-test($p < 0.05$) for comparing between groups, SQ responds quicker than WOZ. The rest of the comparison is not significant. Nevertheless, it shows some preferences of participants. PROPOSAL and WOZ seem to perform better in Q1 and Q2 but do not have a big difference. Participants seem to regard PROPOSAL and

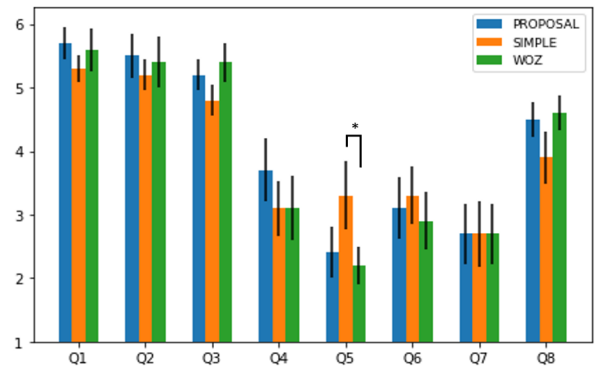

図 5: Questionnaire Result (score of mean and standard error of mean)

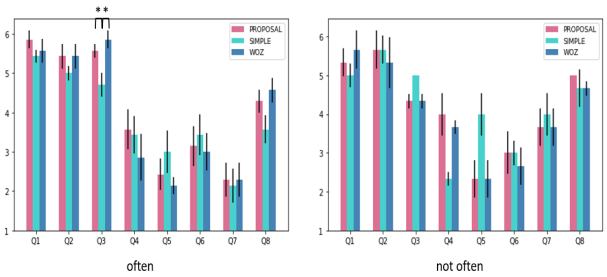

図 6: Attribute Result

WOZ to be more in time than SIMPLE. PROPOSAL seems to respond later. And SIMPLE may be quicker and bothering than the other two. The naturalness may be higher in PROPOSAL and WOZ. The impression result of proposed model and WOZ are similar in most questions except Q4.

In our interview for user feedback, we find there are different impression and user behavior between participants who often go to exhibition(PO) or not(PN) as figure 6. For further analysis, we separate the data and have interesting finding. PO seems to prefer PROPOSAL and WOZ more, while PN may prefer SIMPLE more. PO regard PROPOSAL and WOZ more in time than SIMPLE($p < 0.05$). In contrast, PN seems to prefer SIMPLE more. We also asked PN and they show more desire to have interaction with the agent, and they may expect a more immediate response. The others who expect to visit exhibition guiding prefer observative agent behavior.

Most of our participants have no or less experience in using VR devices. In addition, we didn't inform them of the opening way for agent's introduction behavior. However, it is easy for the participants to get assistance base on agent's perception. Furthermore, some participants regard it as attractive to have exhibition guidance experience in VR and also convenient. However, the reacting time is of the agent is limited to the values of parameters and not adjustable, which makes few participants feel less in time. In addition, the agent is not designed to avoid walking in front of the player that sometimes disturbs participants. Besides our interaction pattern, participants also expect some more interaction with the agent.

6. Conclusion

The research proposed a model for agents to assist players on their initiative by perception of gaze. It allows players to experience guiding tours without direct input toward agents, which is often necessary for human-agent interaction. We create a model for accumulating perception of visual attention and use WoZ approach with optimization to distinguish reacting time and realize proper values for parameters. The model is implemented in VR experiment and has 10 participants to join. It makes people who often go to an exhibition feel more in time.

The problem of our current model is that reacting time is limited base on the optimized parameter values. It is not customized for players and not all players have the preference using our calculation of threshold. Therefore, it is necessary to design a way to simply adjust the values for the individual difference, which may be learned by agent before or during the interacting phase. Also, the interaction mode can be increased to enhance interactivity.

参考文献

[1] Admoni, Henny and Srinivasa, Siddhartha:Predicting user intent through eye gaze for shared autonomy, 2016 AAAI Fall Symposium Series, 2016.

[2] Peters, Christopher and Pelachaud, Catherine and Bevacqua, Elisabetta and Mancini, Maurizio and Poggi, Isabella：A model of attention and interest using gaze behavior, International Workshop on Intelligent Virtual Agents, Springer, pp. 229–240, 2005.

[3] Kevin, Stevanus and Pai, Yun Suen and Kunze, Kai：Virtual gaze: exploring use of gaze as rich interaction method with virtual agent in interactive virtual reality content, Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, pp.1–2, 2018.

[4] 小栗 賢章, 三武 裕玄, 杉森 健, 佐藤 裕仁, 長谷川 晶一(東京工業大学)：多彩で魅力的な触れ合いのできる自律キャラクタ VGent と体験用オンライン VR 環境, 第 25 回日本バーチャルリアリティ学会大会, 2020.

[5] 佐藤裕仁, 三武裕玄, 杉森健, 長谷川晶一：VGentEditor: 操作部位と空間目標点を動作表現として用いたインタラクティブキャラクタの動作生成, 日本バーチャルリアリティ学会大会論文集, 日本バーチャルリアリティ学会, Vol. セッション D-07, 2019.