Virtual Assistant: Enhancing Content Acquisition by Eliciting Information from Humans

Motoyuki Ozeki, Shunichi Maeda, Kanako Obata, and Yuichi Nakamura

Academic Center for Computing and Media Studies, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto, Japan

{ozeki,yuichi}@media.kyoto-u.ac.jp

Abstract: In this paper, we propose the "Virtual Assistant," a novel framework for supporting knowledge capturing in videos. The Virtual Assistant is an artificial agent that simulates a human assistant shown in TV programs and prompts users to provide feedback by asking questions. This framework ensures that sufficient information is provided in the captured content while users interact in a natural and enjoyable way with the agent. We developed a prototype agent based on a chatbot-like approach and applied it to a daily cooking scene. Experimental results demonstrate the potential of the Virtual Assistant framework, as it allows a person to provide feedback easily with few interruptions and elicits a variety of useful information.

Keywords: Semantic ambient media, Embodied agent, Video production, Cooking

1. Introduction

A large number of cameras are now installed in our living environment, for instance, in our homes, our offices, and even our clothes, for the purposes of communication, security, or recording daily life [1][7][9][16][18]. The massive amounts of data obtained from these devices are potential sources of informative content such as education materials or as logs of our daily lives. For example, videos of office work are useful records of the work done and the way it was done, a picture of every meal is useful for our health and weight control, and videos of cooking or DIY become good instruction manuals for children or beginners. Historically, TV or movies have played an important role in providing videos as knowledge, e.g., educational programs and cooking shows. We expect that those media technologies automate those works and extend the application fields.

However, data from ubiquitous cameras and sensors often lack semantic information, e.g., what a person is doing and why, what is important, or where attention should be directed. Such semantic information is particularly important

and often difficult to obtain if a person is passively observed by ubiquitous cameras.

In addition, these types of videos are often poorly organized and not enjoyable compared with TV programs or movies. Although these problems can be solved if we employ camera operators, directors, annotation services, and so forth, we cannot afford these costs for ubiquitous content acquisition.

To cope with this problem, we propose a novel framework, the "Virtual Assistant," which employs an embodied agent with functions similar to a human assistant in TV programs. In TV programs such as cooking shows, one or more persons who help the main performer or instructor, hereafter called "human assistants," often come onto the stage. The human assistant helps the main performer to explain what he or she is doing, carries out instructions, etc. If we can achieve such functions through media technology, it could greatly reduce the disadvantages of ubiquitous video capture.

For this research, we first consider a cooking scene as a target, because the potential content that should be given is clear, and it is easy to evaluate whether appropriate explanations are given. Cooking records can be used in various ways, such as instructions for children, recipe exchanges among friends, or cooking memos for oneself. Moreover, we can see many cooking shows on TV, and the behavior of human assistants in these programs can be good examples for the Virtual Assistant framework.

Through this research, we explore the possibilities of an artificial agent that draws essential information from humans. In addition, we demonstrate that the Virtual Assistant facilitates instructors and elicits essential information in a kitchen where video cameras and other sensors are installed. Thus, the contribution of this research is that it clarifies the ability of the Virtual Assistant through actual experiments in such a ubiquitous/pervasive environment.

The remainder of this paper is organized as follows. In Section 2, we describe the functions of the Virtual Assistant, show examples of human assistant actions shown on TV, and mention related works. We then briefly mention the interaction design in Section 3, introduce our prototype system in Section 4 and demonstrate the Virtual Assistant through some experimental results in Section 5. Finally, we discuss the experimental results and the potential of our research regarding improvement of ambient media experiments in Section 6.

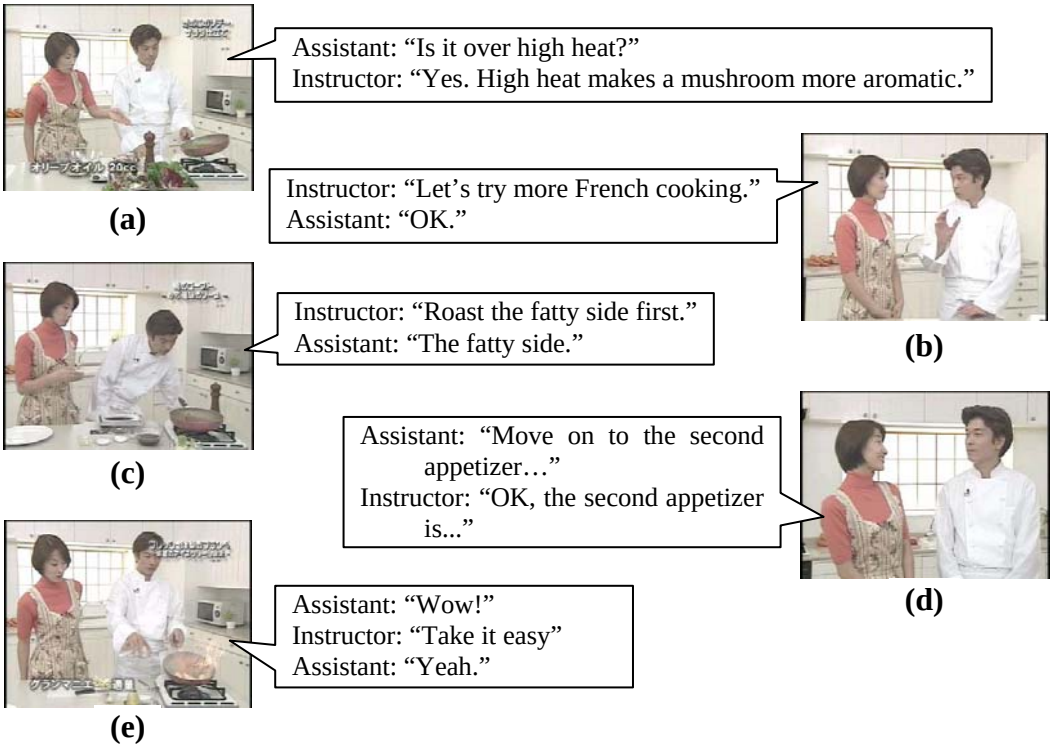

Figure 1: Example behaviors of an instructor and a human assistant in a cooking show.

2. Virtual Assistant

2.1 Functions of an Assistant

Figure 1 shows example behaviors of an instructor and a human assistant in a cooking show. We can easily understand that the functions of a human assistant include (1) adjusting the amount of information in the content, (2) clarifying and sharing the focus of attention, and (3) adjusting the pace and atmosphere of content delivery. These three points are organized as follows:

(1) Adjusting the amount of information

It is often difficult for an instructor alone to organize what should be explained. A human assistant helps by asking questions, adding comments, and providing other reactions to the instructor’s behavior. In Figure 1 (a), the assistant asked about the heat and the instructor naturally added an explanation. This function can be also achieved by an assistant’s expressions or nonverbal behaviors that show interest or curiosity. In Figure 1 (b), the assistant’s response makes it easier for the instructor to talk to the audience. The assistant is a member of the potential

audience and makes a response as its representative. Moreover, such behaviors also help the audience understand the content.

(2) Clarifying and sharing the focus of attention

Attention should be directed to the correct point and shared between an instructor and the audience. A human assistant helps the instructor by expressing interest at important points [Figure 1 (a)] or by repeating important words or phrases used by the instructor [Figure 1 (c)]. The assistant in Figure 1 (d) is not only explaining, but directing the audience's attention to the right point.

(3) Adjusting the pace and atmosphere

Content needs to have suitable pacing to allow the audience enough time to understand without being boring. Greetings or leading to the next section by a human assistant reduces the instructor's burden when the person starts talking [Figure 1 (d)]. A joke, exclamation, or even shriek from an assistant releases the tension and makes the presentation more fun [Figure 1 (e)]. Nodding or repeating important phrases adjusts the pace of speech [Figure 1 (b)].

2.2 Survey of TV Programs

We examined TV programs and counted the human assistants' behaviors in the videos. The TV programs are a cooking show[1] (15.5 min) and a handicraft show for children (13 min).

We first describe typical information appearing in cooking instruction videos. Before starting an actual cooking process, an instructor explains the next process and also occasionally explains the reason the operation is necessary or mentions the ingredients of the food. While actual demonstrations, detailed recipes, methods, information about cookware and ingredients, and tricks are also provided. The quantity or degree of seasoning, heat, and cooking time are also essential information. Much information is elicited from an instructor by an assistant.

Table 1 shows the number of occurrences of typical assistant behaviors in the TV programs. One of the most significant features is that the assistants perform

[1] This cooking video was produced by a professional video production company as a copyright-free sample, and has not been broadcast.

their actions frequently in a short period. In particular, a nod/repeat-phrase type of behavior occurred the most. We consider that this allows the assistant to draw information from the instructor without disturbing his operations and explanations. Other types of behavior also frequently appear, e.g., one every minute or two, and this frequency is also significant considering that each utterance requires a certain amount of time. From this result, we can see that the above functions of a human assistant are frequent and dominant in typical TV programs.

Table 1: Number of occurrences of assistant functions in TV programs. “Atmosphere” means a type of behavior that releases tension, such as a greeting, joke, etc.

TV program	Question	Nod/Repeat-phrase	Additional explanation	Answer	Exhibit interest	Atmosphere
Cooking show	14	122	15	0	7	13
Handicraft show	14	62	4	8	6	20

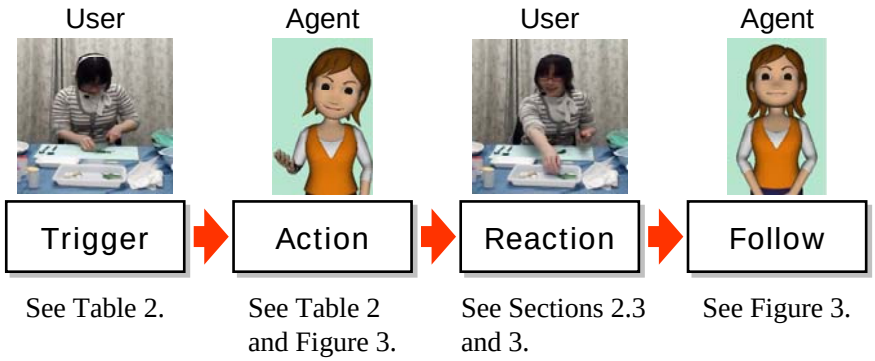

Figure 2: Interaction flow.

2.3 Virtual Assistant Design

To develop an artificial assistant that enables natural and informative interactions, we need a complete AI system with natural language processing, speech recognition, image recognition, etc. Our idea, however, is not heavily dependent on such completeness that is difficult to achieve now.

A possible alternative approach is a chatbot. Chatbot systems often show good performance in drawing information from users and maintaining conversations. From this viewpoint, we first attempted to build a base using a chatbot-like approach, and then gradually add smarter behaviors by introducing AI and media techniques.

In this research, we focus on two effects of the Virtual Assistant: “being present in front of a user” and “asking questions.” For the former, we provided an embodied CG agent. The agent randomly moves even when it is not asking, and this gives the feeling that it is autonomous and waiting for possible interactions. For the latter point, we designed event-based agent behaviors. These behaviors are triggered by typical events in cooking. The system continuously recognizes certain human behaviors and cooking states, and gives pre-determined actions when the system detects trigger events. This approach allows a user to behave naturally compared with a scenario-based approach, in which interactions are scheduled in a detailed scenario.

Our approach is inspired by earlier studies on human-agent interaction [6][10]. These studies proved that artificial agents and their behaviors, such as giving responses, performed well at activating conversations, although these agents have a simple and limited ability for interactions. The important difference between our research and the previous studies is that we focus on content acquisition, which has not been a target of study regarding the usage of artificial agents. Our research originally verified a new possibility for human–agent interaction.

As for the application to cooking, earlier studies dealt with smart kitchens [2][4][5][8][11][15][17]. These studies aimed at cooking support that included event recognition in cooking situations and giving assistance appropriate to the situation. Their methods of object recognition, event detection, and situation recognition are good references to event detection in our approach. Their approaches are, however, mostly based on scenarios that are descriptions of possible event occurrences. In contrast, we adopted an event-based approach, as ambient media should be able to handle events that are not planned beforehand. Our event-based approach deals with the problems in a different way, and shows the possibility of a chatbot-like approach. However, it occasionally causes semantically strange interactions, and we need further study, including investigation of the effective use of scenarios.

3. Implementation of the Interaction

To implement the Virtual Assistant with the event-based approach, we simplify human-agent interactions by considering the following four steps: trigger, action, reaction, and follow (see Figure 2). The process is as follows: If a user’s behavior

meets a certain condition (trigger), an agent asks the user a question (action), the user answers the question (reaction), and the agent responds to the answer (follow).

4. System

Figure 4 shows an overview of our system. Each module is briefly explained below.

Table 2 shows all agent actions used in the experiments. Eight examples are shown in Figure 3. We employed a computer graphics character as a Virtual Assistant.

Each agent action consists of three components: facial expression, motion, and vocalization. For the experiments, we prepared nine categories for the agent actions. Each action category has one or two action(s). An interaction module (described below) chooses an action category, and an action is then randomly chosen from the actions in the category. We expect that random action selection makes the agent's behavior appear natural and autonomous.

4. System

Figure 4 shows an overview of our system. Each module is briefly explained below.

Table 2 also shows some trigger conditions. The action categories and trigger conditions have many-to-many correspondences. Each action category is activated by 1–5 types of triggers, and each trigger is associated with 1–3 types of action categories.

The special type of agent action "do nothing" is invoked at a constant rate. Although the system has a number of other rules, e.g., no trigger is accepted for a certain time after the previous interaction, we omit the details in this paper. A user's reaction is detected by his or her speech.

We prepared three "follow" actions of the Virtual Assistant: convinced, impressed, and perplexed (see Figure 3). If the user says something after the agent's question, the agent responds with a convinced or impressed reaction. If the

user does not answer the agent's question, the agent displays a perplexed expression.

In case no trigger occurs for a certain length of time, the agent displays an "idling" behavior at irregular intervals. We prepared three idling actions of the Virtual Assistant: blinking, nodding, and looking at the table (see Figure 3).

4. System

Figure 4 shows an overview of our system. Each module is briefly explained below.

Table 2: Outline of actions for the Virtual Assistant. Each action category has two types of expressions, obtained by combining facial expression, motion, and vocalization (except "know-how"). Each action category is activated by 1–5 triggers, examples of which are shown in the table.

Action	Facial Expression	Motion	Vocalization
Ask the name of an object or state	Raise eyebrows Smile	Point Tilt the head and point	What is it? I wonder what it is …
Trigger example: User does something for over 5 s without speaking for over 7 s.			
Ask about a situation or condition	Smile	Tilt the head	What are you doing?
		Tilt the head and extend the hand	I wonder what you are doing …
Trigger example: User holds object(s) for over 5 s without speaking for over 7 s.			
Ask about a procedure or method	Raise eyebrows	Tilt the head	How are you doing?
	Smile	Tilt the head and extend the hand	I wonder how you are doing …
Trigger example: User does not hold object(s) for over 5 s without speaking for over 7 s.			
Ask the next step	Raise eyebrows	Tilt the head	How are you doing next?
	Smile	Tilt the head	What are you doing next?
Trigger example: User does not hold object(s) for over 5 s without speaking for over 7 s.			
Ask whether a task is finished	Raise eyebrows	Nod and point	Have you finished?
	Smile	Tilt the head	Are you done?
Trigger example: User puts a cooking utensil on the table.			
Ask about a quantities or degrees	Smile	Tilt the head	How much?
	Smile	Tilt the head	I wonder how much …
Trigger example: User finishes speaking within 1 s after putting a cooking utensil on the table.			

Ask about replacing goods or methods	Smile	Tilt the head	Any other ideas?
	Smile	Tilt the head	What if you cannot do it this way?
Trigger example: User puts seasonings on the table without speaking for over 7 s.			
Ask some know-how	Raise eyebrows	Tilt the head and extend the hand	Do you have any know-how?
Trigger example: User does something for over 5 s without speaking for over 7s.			
Ask a reason for the user's actions	Raise eyebrows	Tilt the head and extend the hand	Why?
	Smile	Tilt the head	I wonder why …
Trigger example: User finishes speaking within 1 s after putting an object on the table.			

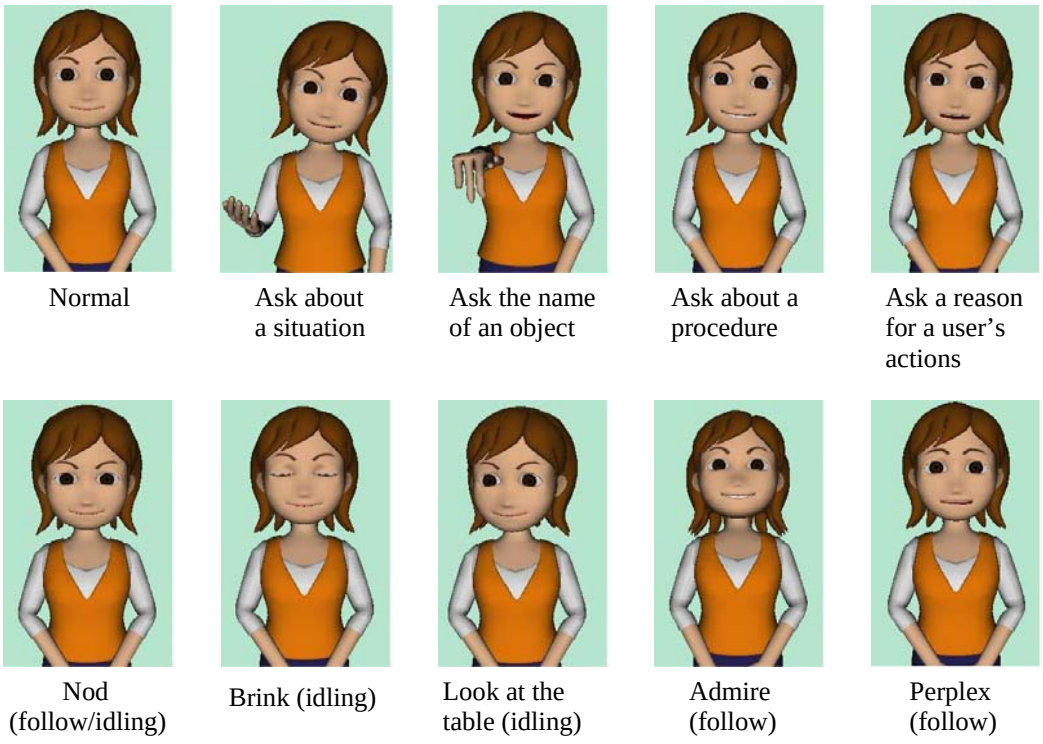

Figure 3: Examples of Virtual Assistant behaviors. The upper figures are actions. The lower figures are follow and idling actions.

4.1 Recognition Module

The recognition module recognizes the typical states of cooking and the user's behavior. This module sends information from the detected triggers to the interaction module. For the experiments, we implemented three functions, which recognize a user operating something with the hands (*operation*), holding an object (*hold*), and speaking (*speech*). By combining these three types of

information, triggers are detected. For example, a user finished talking within 1 second after picking up seasonings.

The *operation* situation is detected from the hand positions. If either one or both hands are in a predetermined workspace, the system considers the situation as *operation*. To detect the position of the hands, two cameras are used, which are attached to the ceiling and wall. First, skin color regions are extracted based on the Mahalanobis distance from our skin color model. The largest and second largest regions are regarded as hand region(s) if their areas are larger than the threshold. The horizontal positions of the hand regions are calculated using the ceiling camera, and the vertical positions using the wall camera.

The *hold* situation is detected by the following steps. Colored markers are attached to objects, and the objects' names and positions are obtained from the colors, sizes, and shapes of their markers. We used three cooking utensils and four seasonings for the experiments. When a user holds an object, the marker on it is occluded by the hand. The module detects that the marker has disappeared and considers the object as being picked up. The *speech* situation is detected by the speech recognition software "Julius." Our system considers only whether or not the user is speaking, and does not consider the spoken words.

Although they are ad hoc methods, they worked well in the experiments. Refinement of the processes should be studied in the future.

4.2 Interaction Module

The interaction module checks whether the condition of each agent action is satisfied: (1) receives triggers from the recognition module, (2) chooses the agent action categories according to the detected trigger, (3) randomly selects an action in the action categories, and (4) sends the name of the selected action to the agent control module.

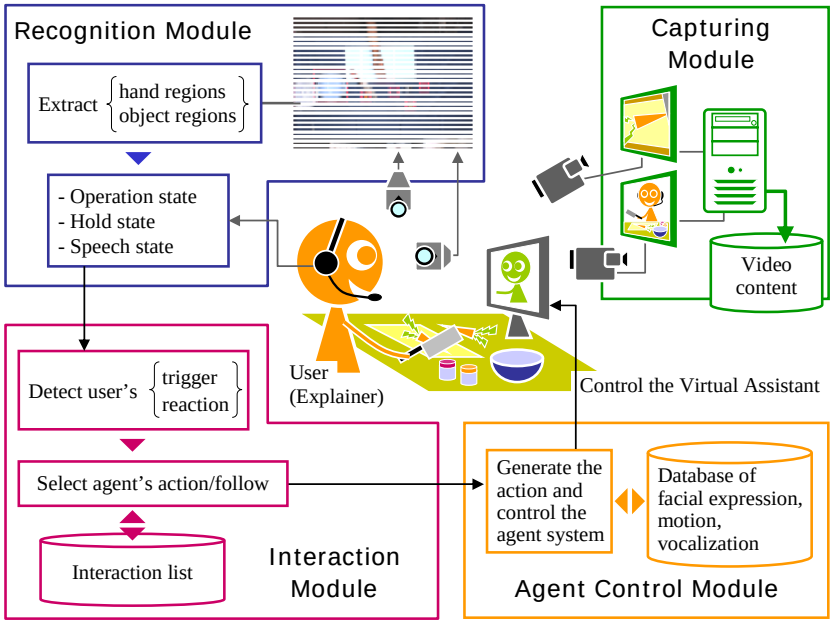

Figure 4: A system overview.

This module switches to a reaction waiting mode after sending the action to the agent control module. In this mode, the module waits for a reaction from the user, e.g., speech, for a certain length of time. If the user begins to speak during this time, the interaction module randomly selects a follow behavior for the agent and sends the behavior name to the agent control module. If the user does not speak during this time, the module sends the behavior name "perplexed" to the agent control module. The interaction module then returns to the trigger waiting mode.

4.3 Agent Control Module

The agent control module generates facial expressions, motion, and vocalization in the agent according to the name of the action received from the interaction module. Facial expressions and motion are created beforehand using CG software (LightWave9, NewTek, Inc.). Voices are also recorded in advance, and are played synchronously as the facial expression changes.

4.4 Capturing Module

The capturing module takes a medium shot of the user and a close-up shot of the workspace. An example is shown in Figure 4, in which the figure of the Virtual Assistant is overlaid in the bottom-right corner as a picture-in-picture signal. Details on video capturing and editing techniques should be studied in the future. Some studies on automatic video production have been reported [3][12][13][14].

5. Experiments

5.1 Purpose and Procedure

We conducted experiments to verify the following points for our Virtual Assistant framework. The association between each of the evaluation items and the assistant's functions (1: amount of information, 2: focus of attention, 3: pace/atmosphere) mentioned in Section 2.1 is given in parentheses.

Evaluation item 1 (EI-1): Whether the content with the Virtual Assistant has enough information, i.e., whether the agent helps a user to provide sufficient feedback (mainly related to function-1).

Evaluation item 2 (EI-2): Whether the context or timing of the Virtual Assistant's questions is appropriate (related to all functions).

Evaluation item 3 (EI-3): Whether the Virtual Assistant provides an environment where the user feels relaxed and encouraged to provide feedback (mainly related to function-3 and function-1).

For comparison, we conducted the experiments under the following three conditions:

No agent: Subjects cook and provide explanations without the Virtual Assistant.

Automated agent: The Virtual Assistant is automatically controlled as mentioned in Section 4.

WOZ agent: The Virtual Assistant is manually controlled by a human operator (Wizard of Oz method). The set of agent behaviors is the same as that of the automated agent.

Comparing no agent with the others demonstrates the efficacy of the Virtual Assistant. Comparing the automated agent with the WOZ agent checks the capabilities of our automated method.

The experimental procedure is as follows. We used eleven subjects (graduate students) as instructors and divided them into two groups: "beginners" (six persons), who rarely cook, and "skilled" (five persons), who often cook without help. No professional cooking instructor was included.

We asked each subject to make a gyoza (Chinese dumpling) and provide explanations in the above three conditions. We prepared a written recipe and gave a copy to each beginner a few days before the experiment. In contrast, the skilled

subjects read the recipe around two minutes before the experiment, because we presume that they already know the rough process, and too much detailed instruction makes them uneasy in worrying about missing details. After each trial, the subjects answered the questionnaire shown in Table 6 (EI-3). After all trials, the explanations in the videos were classified and counted [EI-1 (1)]. We gathered six subjects as the audience. The audience then evaluated the content and timing of the agent's questions (EI-2), and checked whether sufficient information was provided [EI-1 (2)].

In the rest of this section, we report the results of evaluations 1 to 3. In the tables showing the results, we use shortened forms as follows: "B" and "S" indicate beginners and skilled cooks, respectively; "None," "Auto," and "WOZ" mean no agent, automated agent, and WOZ agent, respectively.

In this section, we first mention each of the experimental results, and then discuss them all.

5.2 [EI-1] Amount of Information in the Content

We first confirmed whether or not the amount of information in the content was affected by the Virtual Assistant.

(1) Frequency of Explanations

Table 3: Number of explanations provided. The table shows the average number of explanations provided by all subjects.

Explanation Category	**None**		**Auto**		**WOZ**	
	B	S	B	S	B	S
Procedure/Method	9	12	5	10	3	11
State/Condition	12	22	14	7	11	7
Quantity/Degree	30	72	17	30	12	25
Know-how/Reason	40	52	28	32	19	22

Table 3 shows the frequency of the subjects' explanations. These explanations are categorized into four typical types: procedure/method, state/condition, quantity/degree, and know-how/reason.

Regarding the procedure/method and state/condition, there is no significant difference between using and not using the Virtual Assistant in both the beginners and the skilled subjects. This is because it is fairly easy for users to explain what they are doing while they are doing it. Meanwhile, regarding the quantity/degree

and the know-how/reason, we can see that the agent successfully drew important information from the subjects by asking questions.

The significant difference between beginners and skilled subjects is the frequency of feedback for know-how/reason. This difference arises from the beginners' lack of knowledge. The beginners were often unable to find a good answer to these types of questions and occasionally answered "Nothing." or ignored the question. On the other hand, the skilled subjects answered well by flexibly interpreting ambiguous questions.

(2) Subjective Evaluation by Audience

Table 4 shows the number of items that were regarded as lacking information by the audience. Each number is a sum of the results for all the subjects, and the average number of items lacking information for each subject is roughly one-half to two-thirds. For the evaluation, six videos of around five minutes were taken from two videos of beginners and two of experts under the no agent, automated agent, and WOZ agent conditions.

Table 4: Number of items lacking information as judged by the audience. The table shows the total number of results obtained from the entire audience.

Items lacking in information	**None**		**Auto**		**WOZ**	
	B	S	B	S	B	S
Procedure/Method	9	12	5	10	3	11
State/Condition	12	22	14	7	11	7
Quantity/Degree	30	72	17	30	12	25
Know-how/Reason	40	52	28	32	19	22
Total	91	158	64	79	45	65

We can see that the number of items lacking information is greatly decreased using the Virtual Assistant. This demonstrates the good potential of the Virtual Assistant for eliciting a wide variety of information from individuals, especially the quantity/degree and know-how/reason information. In contrast, the results for procedure/method information are not affected as much by the Virtual Assistant because this type of explanation is relatively easy to speak even in a solo instruction.

On the other hand, the results obtained from the skilled subjects were worse than those from the beginners because the former tended to provide a rougher, less detailed explanation, especially regarding the procedure/method and quantity/degree, and beginners in cooking wanted more accurate information.

This causes a content–audience mismatch. We may need to design the Virtual Assistant so that it is adjustable to the audience.

5.3 [EI-2] Evaluations of Content and Timing of the Agent's Questions

Table 5: Evaluation of content and timing of agent's questions by the audience. The table shows the percentages of the agent's questions that the audience considered adequate.

	Auto		**WOZ**	
	B	S	B	S
Adequate content	61.5%	57.0%	79.2%	94.0%
Adequate timing	59.4%	57.0%	81.9%	82.1%
Both timing and content adequate	39.6%	39.5%	65.3%	78.6%
Number of questions	96	114	72	84

Table 5 shows the evaluation results of the content and timing of the agent's questions. In the case of the WOZ agent, from 65% to 80% of the agent's questions were evaluated as "Both timing and content adequate." This result suggests that an event-based approach using combinations of a few fixed agent behaviors can provide adequate help to an instructor even without a detailed scenario for the instruction.

On the other hand, the scores with the automated agent were worse than these with the WOZ agent. The automated agent occasionally asked similar questions two or more times and talked (typically back-channeling) while the instructor was speaking. These interactions made the score significantly lower than in the WOZ agent case. Improvements can be achieved by more advanced speech recognition techniques, which clarify the type of explanation being provided, e.g., method, reason, know-how, and distinguishing the intake of breath and the end of a sentence more accurately.

5.4 [EI-3] Subjective Evaluations by Instructors

Table 6: Subjective evaluation of the subjects (=instructors). This table shows the average scores of all the subjects. (1: Strongly disagree—3: Neutral—5: Strongly agree)

Questions	**None**		**Auto**		**WOZ**	
	B	S	B	S	B	S
You were able to concentrate on the cooking	3.5	4.4	3.2	3.4	3.5	3.8
You were able to enjoy the cooking	2.3	3.2	3.7	3.0	3.2	3.2
You were able to pleasantly explain your actions	2.0	2.2	3.2	2.6	3.3	3.4
You were able to be aware of the audience	2.8	2.8	3.2	2.8	2.8	4.0
You were able to provide convincing explanations	2.3	2.8	3.0	2.6	3.3	3.2
The agent was friendly to you	-	-	2.7	2.8	3.0	3.2

The agent's actions were natural	-	-	2.3	2.2	3.0	2.6
The timings of the agent's actions were good	-	-	2.5	2.4	2.8	2.6
The frequency of the agent's actions was good	-	-	2.7	2.4	3.2	3.4

Table 6 shows the questionnaire results from those subjects who provided feedback. Ratings range from 1 to 5 (1: Strongly disagree—3: Neutral—5: Strongly agree).

We first mention the upper items, from the first item to "to provide convincing explanations." Most of the results for the WOZ agent show higher marks than those of the no-agent situation. This suggests that the Virtual Assistant can enhance the instructor's explanation without disturbing the work if the agent could be ideally controlled. The automated agent received much lower marks than the WOZ agent. The beginners, however, occasionally seemed to enjoy the agent's strange behaviors and gave a higher assessment to the automated agent.

The results obtained from the skilled subjects differ from those of the beginners. The skilled subjects seem to have evaluated the agent rigorously. In particular, regarding "to concentrate on the work," the score for no agent was much higher than those of the other types of agent. On the other hand, in the above evaluations, lack of information is often reported more for the skilled subjects. It shows that skilled subjects tend to skip explanations that the audience may require. We can observe certain trade-off between the amount of information and users' feeling of disturbed. In addition, the difference between the automated and WOZ agents was relatively large compared with the result from the beginners. This is because the skilled subjects paid more attention to interactions with the agent, and the quality of the interactions was more crucial for them than for the beginners.

We next mention the lower items, from "The agent was friendly to you" to the last item. The Virtual Assistant obtained low marks on all these items, especially for the automated agent. The timing, frequency, and content of the agent's question should be improved, as mentioned in Section 5.3. Friendliness depends on the way of speaking or the appearance, and a human-like appearance may be inappropriate on this point because we expect too much ability from the agent. We are now making another agent character, a dog.

6. Towards Ambient Media

We currently focus on typical situations such as cooking or DIY, as people surely want to maintain records of their activities, and we can expect their cooperation with the recording system. We conjecture that the Virtual Assistant extends content production to a more ambient approach for taking our activity records.

Careful considerations are, however, necessary. The following are drawbacks to our approach.

- It is questionable whether such agents are widely accepted by ordinary people. Agents can be noisy, annoying, or obtrusive. Such systems will eventually be powered off.
- It is questionable whether we need anthropomorphous agents. A simple notice, such as a written sentence, might be sufficient.
- Artificial agents may cost too much.

We do not have clear answers to these drawbacks. However, our opinion is as follows:

- Agent designs can provide partial solutions. We can find it difficult to ignore human-like agents when they speak to us. However, we can do as we like if they make a slight bow or smile at us. Well-designed nonverbal behaviors would have powerful functions to draw our reactions, while they can also be easily ignored.
- Observing users and estimating to what extent or how they can allow the agent's intervention is an interesting research topic. The topic is closely related to general problems in ambient media and ubiquitous computing. Human behaviors such as head motion, gazing, and attitudes in speaking will be good clues for estimating it.

We hope that these points will be clarified in the near future through intensive research on agent approach, ambient media, ubiquitous computing, and related fields.

7. Conclusion

In this paper, we have proposed the Virtual Assistant framework that enhances communications between humans and ambient media. We have developed a

prototype agent based on a chatbot-like approach and demonstrated its potential in real-life explanations regarding cooking. Although our experiments are preliminary and limited to cooking scenes, they show that a simple chatbot-like agent has the ability to elicit information from humans. With few interruptions, the Virtual Assistant helps users to explain what they are doing and lift their face toward a camera, and elicits a variety of information when something should be explained further. We believe that the basic idea of the Virtual Assistant can be used in a wide variety of situations, as the use of questions and answers is common in our everyday behavior.

On the other hand, there is much room for improvement. Agent behaviors are far from satisfactory. The performance of an automated agent is worse than the human-operated case, mainly because the former cannot recognize the context, and thus, cannot adapt its behavior accordingly. More advanced speech recognition and natural language processing will greatly improve this. More accurate image processing will also help the agent to recognize situations. What a user (the person providing commentary) thinks about the agent's unsophisticated behavior and how this user can utilize it is an interesting, open problem. The differences between the beginners and skilled subjects in the experiments also show interesting phenomena. In addition, the Virtual Assistant will be more useful if it can be adjusted to the potential audience, including the commenter, family, beginners, or children. We require numerous additional experiments and further clarification of the underlying in order to determine better what information is necessary for whom.

Through this research, we have explored the possibilities of an artificial agent that draws essential information from humans in a ubiquitous/pervasive environment. We expect that everyone is surrounded by ambient agents, that our experiences are naturally recorded, and that the obtained content has the potential to be used in these environments. In that sense, we have been regarding our Virtual Assistant as an essential component of ambient media.

Reference

1. Abowd G D, Atkeson C, Bobick A, Essa I, MacIntyre B, Mynatt E, T. Starner (2000) Living laboratories: The future computing environments group at the georgia institute of technology.

In: Extended Abstracts of the ACM Conference on Human Factors in Computing Systems, pp 215-216

2. Chi P, Chen J, Chu H, Chen B (2007) Enabling nutrition-aware cooking in a smart kitchen. In: Extended Abstract of Human factors in computing systems, pp 2333-2338
3. Davis M (2003) Active capture: Automatic direction for automatic movies. In: Proc. 11th Annual ACM International Conference on Multimedia, pp 602-603
4. Hamada R, Miura K, Ide I, Satoh S, Sakai S, Tanaka H (2004) Multimedia integration for cooking video indexing. In: Proc. Pacific-Rim Conference on Multimedia, II:657-664
5. Hashimoto A, Mori N, Funatomi T, Yamakata Y, Kakusho K, Minoh M (2008) Smart kitchen: A user centric cooking support system. In: Proc. Information Processing and Management of Uncertainty in Knowledge-Based System, pp 848-854
6. Imai M, (2003) Physical relation and expression: Joint attention for human-robot interaction. IEEE Transaction on Industrial Electronics, 50(4):636-643
7. Ishii H, Wisneski C, Brave S, Dahley A, Gorbet M, Ullmer B, Yarin P (1998) Ambient room: Integrating ambient media with architectural space. In: Proc. Conference on Human Factors in Computing Systems, pp 173-174
8. Ju W, Hurwitz R, Judd T, Lee B (2001) Counteractive: An interactive cookbook for the kitchen counter. In: Extended Abstracts of CHI 2001, pp 269-270
9. Lugmayr A (2007) Ambient, ambient, ambient - what are ambient media? In: TICSP Adjunct Proc. of EuroITV 2007, pp 89-93
10. Ishii R, Miyajima T, Fujita K, Nakano Y. (2006) Avatar's Gaze Control to Facilitate Conversational Turn-taking in Virtual-Space Multi-User Voice Chat System, In: Proc. Int'l Conf. on Intelligent Virtual Agents, Lecture Notes in Computer Science, 4133/2006, Intelligent Virtual Agents, pp 458
11. Nakauchi Y, Fukuda T, Noguchi K, Matsubara T (2005) Intelligent kitchen: Cooking support by lcd and mobile robot with ic-labeled objects. In: Proc. Int'l Conf. on Intelligent Robots and Systems
12. Ozeki M, Nakamura Y, Ohta Y (2005) Automated camerawork for capturing desktop presentations. IEE Proc. on Vision, Image ¥& Signal Processing, 152(4):437-447
13. Pinhanez C, Bobick A (1997) Intelligent studios: Modeling space and action to control tv cameras. Applications of Artificial Intelligence, 11(4):285-306
14. Rougvie M, Olivier P (2007) Dynamic editing methods for interactively adapting cinematographic style. In: TICSP Adjunct Proc. of EuroITV 2007, pp 304-308
15. Siio I, Mima N, Frank I, Ono T, Weintraub H (2004) Making recipes in the kitchen of the future. In: Extended abstracts of Human factors in computing systems, pp 1554-1554
16. Tapia EM, Intille SS, Larson K (2004) Activity recognition in the home setting using simple and ubiquitous sensors. In: Proc. Pervasive 2004, LNCS 3001, pp 158-175
17. Tran QT, Calcaterra G, Mynatt ED (2005) Cook's collage: Deja vu display for a home kitchen. In: Proc. Home-Oriented Informatics and Telematics 2005, pp 15-32
18. Yamazaki T (2005) Ubiquitous home: Real-life testbed for home context-aware service. In: Proc. Tridentcom 2005, pp 54-59

Motoyuki Ozeki received his B.E, M.E. and Ph.D. degrees in engineering from University of Tsukuba, in 2000 and 2005, respectively. He is currently an Assistant Professor at Kyoto University. His research interests are in the areas of human-agent interaction and cognitive science.

Shunichi Maeda received his B.E and M.E. degrees in electronical engineering from Kyoto University, in 2008. He is currently working in Patent Office (KAJI-SUHARA & ASSOCIATES).

Kanako Obata received her B.E. degree in economics from Osaka Prefecture University in2004. She is currently an educational assistant at Kyoto University as since 2004. Her research interests are human-communication and cooking.

Yuichi Nakamura received his BE degree in 1985, his ME and PhD degrees in electronical engineering from Kyoto University in 1987 and 1992, respectively. He worked as assistant professor at University of Tsukuba since 1993 and as associate professor since 1999. He is currently a professor at Kyoto University. His research interests and activities include human-computer interactions, video analysis, and video utilization for knowledge sources.