View Angle Evaluation of A First-person Video to Support An Object-finding Task

Takahiro Ueoka, Tatsuyuki Kawamura, Yasuyuki Kono and Masatsugu Kidode

Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0192, Japan

{taka-ue, kawamura, kono, kidode} @is.naist.jp

Abstract: The aim of this paper is to investigate an effective view angle of "a first-person video" when a user tries "an object-finding task", a task to find a target object in his/her everyday environment. The first-person video means a video recorded by a head-mounted camera of a wearable system worn by the user. The object-finding task can occur when the user forgets where he/she last placed the target object. The first-person video can include some contexts of the event. There are two important contexts required by the user to find the target object; one is "the action" that he/she placed the target object, and the other context is "the location" that the object was placed. The view angle of a lens, which is employed by the camera device, affects the width and the time length of the first-person video. It is important to investigate appropriate range of the angle so that the user can recognize those contexts by watching the first-person video. We conducted experiments that subjects evaluate videos in several width of the angle. We then found that the range of the view angle is from 115 to 125 in degree.

Keywords: view angle evaluation, object-finding support, first-person video, augmented memory.

1 Introduction

In this paper we investigate an effective view angle of "a first-person video" to support a user's "object-finding task" in his/her everyday environment. The first-person video is recorded by a head-mounted camera worn by the user. The first-person video can illustrate contexts of the user's experiences.

We have designed a wearable system to support a user's object-finding task by showing him/her the first-person video. The object-finding task can occur when the user forgets where he/she last placed the target object. It is estimated that a person wastes 150 hours a year in the object-finding tasks [1]. By decreasing the wasted time, a person can utilize the surplus time for other work and leisure, which may enrich his/her everyday life.

In recent years, augmented memory has been investigated extensively in the field of wearable computing [2, 3]. We have proposed a wearable augmented memory system named "*I'm Here!*" to support a user's object-finding task [4]. We have also conducted experiments to evaluate the effectiveness of the first-person video displayed by *I'm Here!* [5]. In those experiment we assumed a user who could not recall anything about an event that he/she last placed a target object on a location. Subjects used the *I'm Here!* system to watch experimental first-person videos. They then performed object-finding tasks in an experimental environment, which corresponded to the videos. We investigated their time efficiencies in experimental object-finding tasks. In the result, we found that *I'm Here!* has the ability to make a user's object-finding tasks efficient when the user watches the output of *I'm Here!*.

Some related works for supporting the object-finding task have been studied. "*Hide and Seek*" [6] is a system that navigates with the frequency of sound how far a target object is placed from a user. Each active tag is attached to each target object to generate the sound. *Hide and Seek*, however, cannot support the object-finding task when the target object has been placed in a location where the user cannot hear the sound from the object. By showing the user a first-person video illustrating the location, *I'm Here!* can support the object-finding task regardless of the position and distance between the

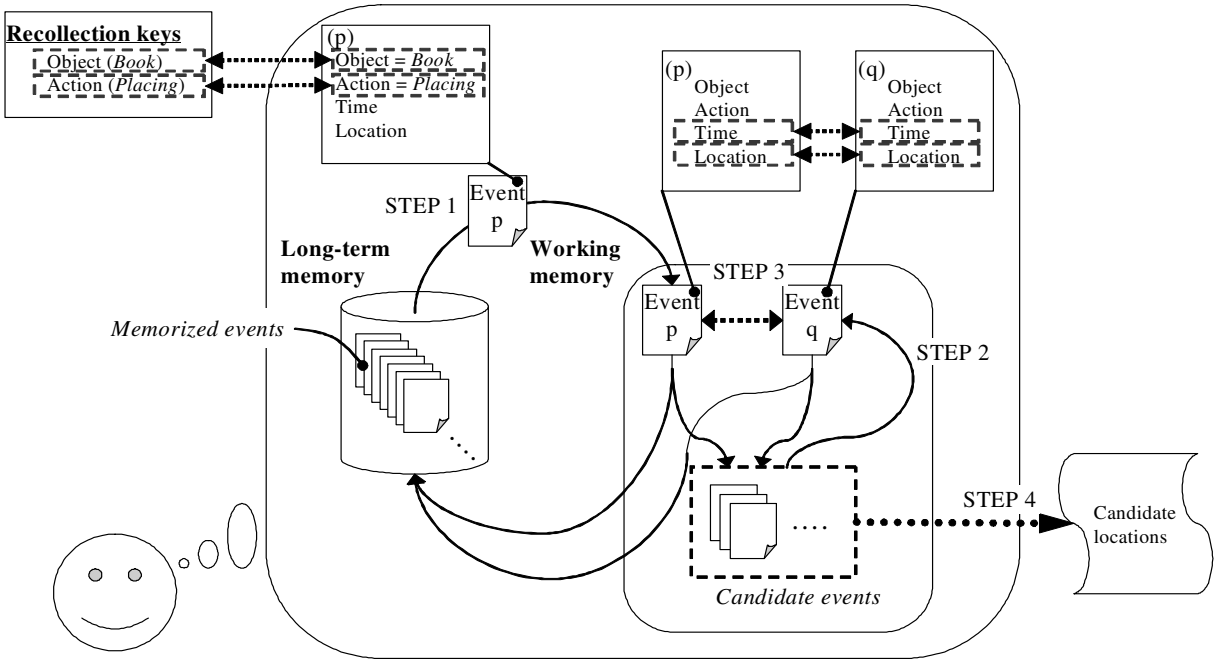

Figure 1: A decision process model of candidate places on an object finding task

user and the target object. "*iFlashBack*" [7] is a kind of wearable video retrieval system to support a user's ability of memorization. *iFlashBack* automatically records a first-person video when the user handles a target object, and then replays the video immediately so that he/she can memorize the scene . Considering, however, the large number of events that the user handles objects in a day, the user's cost for memorization is expected to be large. *I'm Here!* shows the user the first-person video only when he/she wants to recall the location of the target object without requiring a large cost for to the user for memorization efforts..

On the other hand, *I'm Here!* will not fully support the user's object-finding task. For example, *I'm Here!* actually shows the user a first-person video of the event where the user last held the target object. Even though the user placed the object at the end of the event, the video may not always include the moment that the user placed the target object. We suspect that this problem is caused by a view angle limitation of the head-mounted camera employed by the *I'm Here!* system.

In Section 2, we explain how the first-person video with a limited view angle supports a person's memory activities in his/her everyday object-finding tasks. Section 3 discusses the conditions and results of experiments conducted using subjects to investigate the effectiveness of the view angle of the video. Following these results, in Section 4, we discuss how the video should be displayed to improve the efficiency of the object-finding task. In Section 5, we conclude this paper with a future proposal for how to make improvements in the *I'm Here!* system.

2 View angle of a first-person video to support an object finding task

The aim of this section is to describe problems occurring when the object-finding support system shows a user a first-person video to support his/her object-finding task. The problem is that a user's object-finding task can be adversely affected by the first-person video, which has a limitation in its view angle. In this chapter, we first propose a model of human memory activity to solve an object-finding task. With the use of the model, we then explain how the first-person video works on memory activity. Finally, we describe a first-person video where the limitation of its view angle and time length can have adverse effects on the object-finding task.

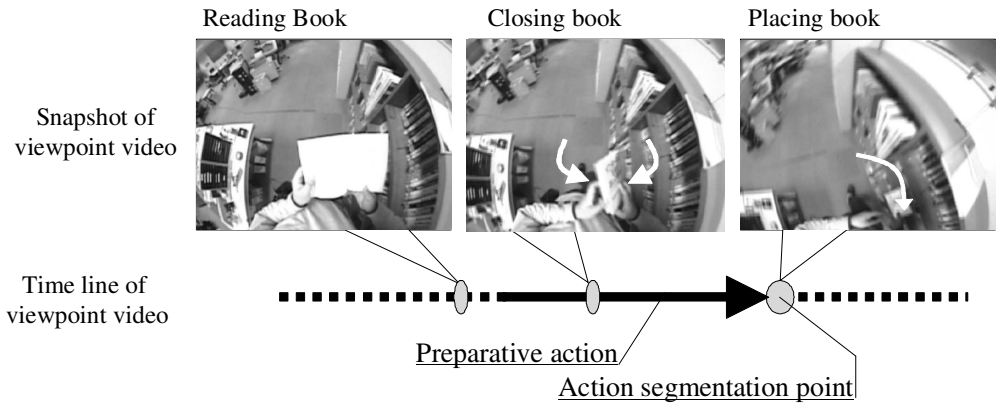

Figure 2: Overview of contexts for perceiving the action of placing a book.

2.1 Model of human memory to find an object

We first design a model of a human memory activity related to an object-finding task. We assume that a person accumulates contexts of an event he/she experienced into a long-term memory through the memorization process. The contexts consist of four elements; object, action, time, and location. For example, the object context includes placing a target object. Features of the target include name, appearance, and texture. Each of these elements has a certainty factor called "clarity". For example, the clarity of a time-element fades with time, or is attenuated by interference with other events.

Next, we explain an ambiguity problem in a time estimation of an event caused by a time-element with less clarity; for example, the night before a person temporarily placed a book at a certain location at 19:00, and then, the same night moved the book to another location at 20:00. If the time-element clarity of the last event has been attenuated, the user might remember that the last event occurred the night before. As a result, the last event and the previous event may interfere with each other, and the user may not be able to decide which event last occurred.

Figure 1 illustrates a model of a person's memory activity to remember where he/she last placed a target object. In the model, human memory consists of a long-term memory and a working memory. The long-term memory accumulates events he/she has been experienced for a long period. The working memory carries out functions of event extraction and event evaluation so as to select required events from the large number of events accumulated in the long-term memory.

STEP 1: the person first determines a target object. He/she then retrieves an event p, which has object-element and action-element according to the target object and the placing action, from the long-term memory. The event *p* is transferred into the working memory. If the event *p* has action-element with less clarity, it is not transferred.

STEP 2: Among events which has been already gathered on the working memory, an event *q* is picked up which has never been compared with the event *p*. If there is no event preliminarily transferred into the working memory, the person carries out STEP 1 again with remaining the event *p* on the working memory.

STEP 3: The person compares the time-elements in the events *p* and *q* in terms of newness. If he/she can decide that *q* is older than *p* with certain clarities of their time-elements, he/she gets *q* back into the long-term memory with remaining *p* on the working memory. If not, he/she leaves *q* on the working memory with getting *p* back into the long-term memory. Only when he/she cannot decide which event is older, he/she leaves both of them on the working memory. Cycles of STEP 2 and STEP 3 are carried out until all gathered events are compared with *p*.

STEP 4: If the person is satisfied with candidate events or if his/her time is running out, he/she

halts a number of repeats from STEP 1 to STEP 3. He/she then extracts candidate locations from location-elements of the candidate events which have been gathered on the working memory.

To carry out the object-finding task effectively, the person has to remember where he/she last placed the target object with a certainty. The event which he/she last placed the target object should have a time-element with large clarity so as to avoid interference with other events. On the other hand, the number of candidate events is desired to be as small as possible.

By the use of the model mentioned above, problems in a person's memory activity for an object-finding task can be classified into following patterns:

(i) He/she cannot recall any event which he/she last placed the target object.
(ii) He/she tend to recall a number of locations on where he/she might place the object.
(iii) He/she cannot recall any location even though he/she remember that he/she placed the object on somewhere.

We call the event which he/she last placed the target object on the location as "correct-event". The problem (i) can occur if the action-element in the correct-event, i.e. the placing action, has less clarity. The correct-event with attenuated action-element cannot be transferred into the working memory in STEP 1. In the result, outputs in STEP 4 cannot include the location-element of the correct-event.

The problem (ii) can occur with less clarity of the time-element in the correct-event. Because of interference of the event p and q in time-element, both events are left on the working memory. In the result, the person outputs a number of candidate locations extracted from candidate events. The problem (iii) can occur if the location-element in the correct-event, which indicates the correct location where the person placed the target object, has less clarity. The correct-event can be gathered on the working memory in STEP 2, but the location-element with less clarity has no information to be reflected in outputs in STEP 4. In the result, the person cannot recall the location where the target object was placed.

2.2 Effectiveness of viewpoint video for human memory

The user can get contexts of a correct-event from a first-person video of the event when the system shows the user the video preliminarily recorded with HMD. The user can get clarity of the time-element in the correct-event accumulated in his/her long-term memory. In the result, the user can find the target object effectively.

The contexts consist of location and action. The location is illustrated by landmarks, and the action is denoted by continuous movement of the hand with the target object as an action. If the user recognizes these contexts, he/she can use them to perceive the action of placing the target object and to identify the location appearing on the video. This action of recollection and location recollection, mentioned in Figure 1, will be reinforced by the results of (1) and (2) below:

(1) Newtson et al. note that human can perceive a continuous action stream as a sequence of clearly segmented action units [8]. When a user watching a viewpoint video perceives the action placing the target object, he/she should recognize any contexts in the video necessary for the perception. We assume that there is a key context called "action segmentation point." We also assume that the action segmentation point accompanies "preparative action," such as a movement of hand with the target object preparing to place. Figure 2 illustrates the overview of them.
(2) When he/she watches a viewpoint video to find a target object, he/she should identify a location appeared in video to compare with his/her memory of the location. Landmarks, such as objects preliminarily set up on a location, are necessary to identify the location. We assume that a person's memory of location, included in an episodic memory, can be represented by combination of position and character of landmarks. With comparing such features of landmarks, he/she can narrow down number of candidate locations. Note that everyday environment consists of a number of locations linking to each other directly or indirectly. Even if a person cannot identify a location appeared in a scene of video, he/she still be able to estimate where the location is with relationship between locations appeared before or after in the video.

If the clarity of the time-element in the correct-event increases, the user will be able to get the following benefits in the model of memory activity:

(a) The action-element in the correct-event will be reinforced, and the user can avoid problem (i) by making STEP 1 well.

(b) Because the clarity of the time-element in the correct-event has been increased, the user can carry the other events back to the long-term memory in STEP 3. In the result, the number of remaining events can be effectively narrowed down to avoid the problem (ii).

(c) The location-element in the correct-event will be reinforced, and the user can recall the correct location in STEP 4 and avoid the problem (iii).

2.3 Limitation of viewpoint video

As we described in the previous section, the system has benefits to basically support the user's object-finding task by showing the first-person video. The first-person video, however, has a limitation in its view angle which causes problems such as when the user is unable to receive the benefits of the video. The user's context recognition load increases because of the limitation. And worse yet, the user can misunderstand the event illustrated in the first-person video.

The amount of contexts illustrated in the first-person video changes with the view angle of a lens employed in a wearable camera. The variation of view angles affects both the "view size" and the "time length" of the video simultaneously. Here we assume three conditions or rules regarding the wearable camera: 1) the system records a first-person video only when the user handles a target object in sight of the camera, 2) an installation condition requires the camera to always keep the same attitude angle against the user's head, and 3) a resolution condition that any first-person video must be displayed in a specific resolution. If the view angle becomes narrower, the target object illustrated in the first-person video becomes larger. However, the target object tends to drop out from the sight in the first-person video. The time length during which the object is illustrated in the video would become shorter. Simultaneously, the possibility to contain contexts of both action segmentation point and preparative action decreases. In such a case, the user would not understand the contexts which may occur. The number of landmarks also decreases even though the user can watch each landmark clearly. On the other hand, if the view angle becomes wide, the target object becomes harder to drop out from the sight of the first-person video, and the time length of the video may become longer. The number of landmarks may increase. However, the size of landmarks and environmental appearances will become smaller and less understandable because the first-person video has only a specific resolution. We assume that the effectiveness of wider view angle is limited by a ceiling.

The conditions on the viewpoint video are inevitably constrained as below:

Limitation of view angle

The maximum view size of the video recorded in the system is defined by the lens employed in the head-mounted camera. Required contexts to reinforce the user's action/location recollection tend to drop out from the limited view size area.

Limitation of time length

The setting of the view size also affects the time length of the first-person video shown to the user. The definition of time length is based on an index of the target object associated with the video. Simply, the end point of the time length is defined as the last frame in which the target object was observed, and the start point is defined as the frame traced back some seconds from the end point. If the view size is too small, the end point will be moved forward. The action segmentation point, mentioned in Figure 2, can run out of the limited time length.

The first-person video with constrained view size and time length can cause the user following problems:

Misunderstanding of the action

The system can show the user a first-person video of a "wrong-event" recorded when the user keeps handling the target object outside of the sight of the first-person video. Because of the limited view angle, the system cannot differentiate between the correct-event and the wrong-event. If the user watches the video of a wrong-event, and if he/she misunderstands the action illustrated in the video as an action of placing, the user tends to replace the event, which is associated with the wrong-event, by a pseudo correct-event which has the pseudo action of placing.

Recognizing confusing action

If the user watches a first-person video which illustrates an event with a confusing action, then the user can pick up the event from long-term memory by using any contexts of the video except the action. However, the user can hardly be sure whether the event is actually a correct-event or not. Even though the event is

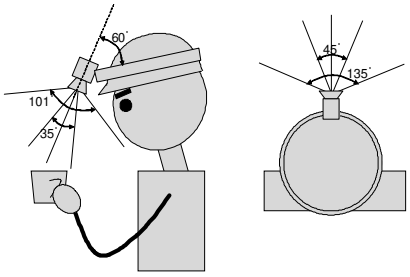

Figure 3: A camera device to record source videos.

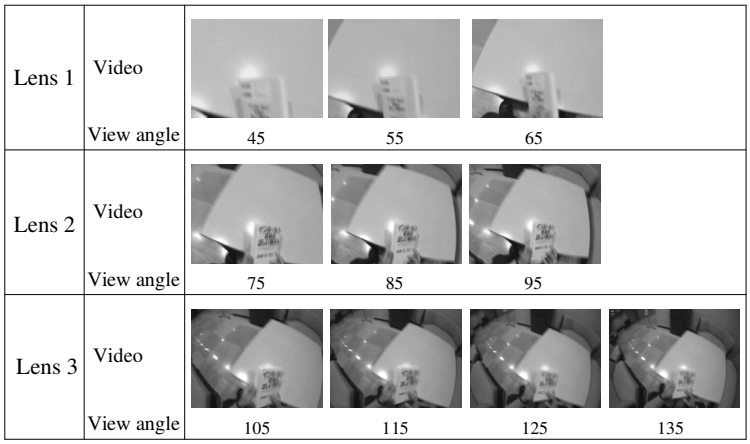

Figure 4: Viewpoint video trimmed in view angle stages.

left on the working memory as a candidate event, the event makes no sense for object-finding support. What is worse, action recognition with less clarity can burden the cognitive capability of the user.

Recognizing confusing location

When the user watches a first-person video illustrating a confusing location, the user can recognize the location as being related to a number of events accumulated in the long-term memory. The number of candidate locations for the user's object-finding task can increase if those events are transferred into the working memory as candidate events.

To avoid those problems, the best-suited conditions of the view angle should be investigated.

3 Experimental evaluation

To analyze a suitable view angle for object-finding support, we performed experiments to investigate what subjects watching virtual first-person videos recognized. The virtual first-person videos were segmented preliminarily with the same methods *I'm Here!* employs. We have discussed how to construct the first-person video suited to supporting a user's object-finding task by analyzing the experimental results.

3.1 Conditions

Fifteen subjects participated in these experiments. All of the subjects were male students of a graduate school who studied information science in a laboratory. In an experimental trial, each subject watched a video displayed on a LCD to answer a questionnaire about the context of the video. The size of the LCD is 16.0 inches.

All experimental videos, which are videos prepared to be shown to each subject, displayed scenes from an experimenter's viewpoint recorded by a head-mounted CCD camera the experimenter wore (Figure 3). The camera was mounted on the front of his head with a 60 degree of depression viewpoint. In this experiment, the resolution of the videos is 220 by 165 pixels. All videos were displayed on the LCD at the same magnification.

The view size conditions of the videos were virtually simulated with trimmed source videos (Figure 4). To establish a wide range of view size conditions, 3 source videos were recorded with 3 types of lenses alternately attached to the camera. The source videos were trimmed in a total of 10 view size stages to simulate view size conditions.

Each source video was recorded when the experimenter selected a positive or negative action (see below) by placing a target object, in this case a book, in a target place. 10 target places in the laboratory where students spent time were selected (Figure 5). A positive action represents an experimenter actually placing the target object in a target location. In contrast, a negative action represents an experimenter not placing the target object in a target location, but first holding on to the object and then bringing it to another location, not designated as a target place. Positive actions were held in 5 places, half of all the places, and negative actions were held in the other half. The experimenter conducted these behaviours naturally while recording videos.

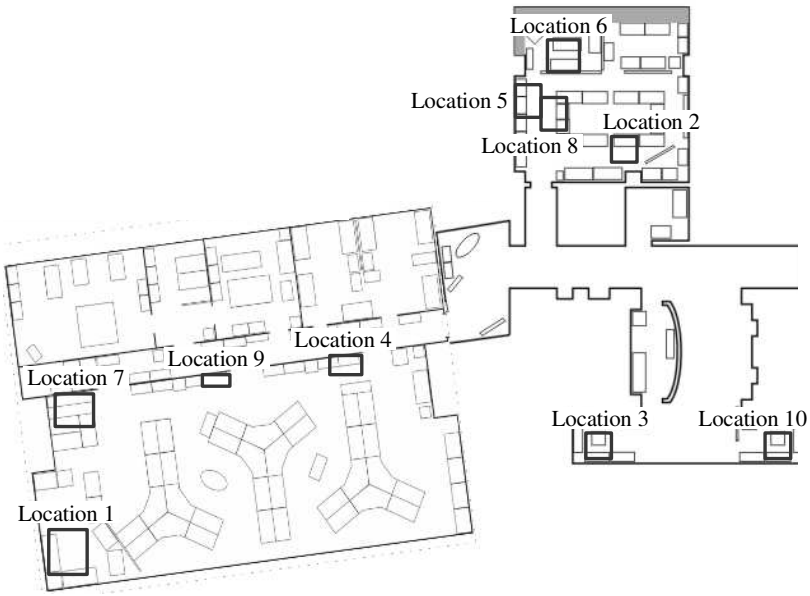

Figure 5: Laboratory environment and target places.

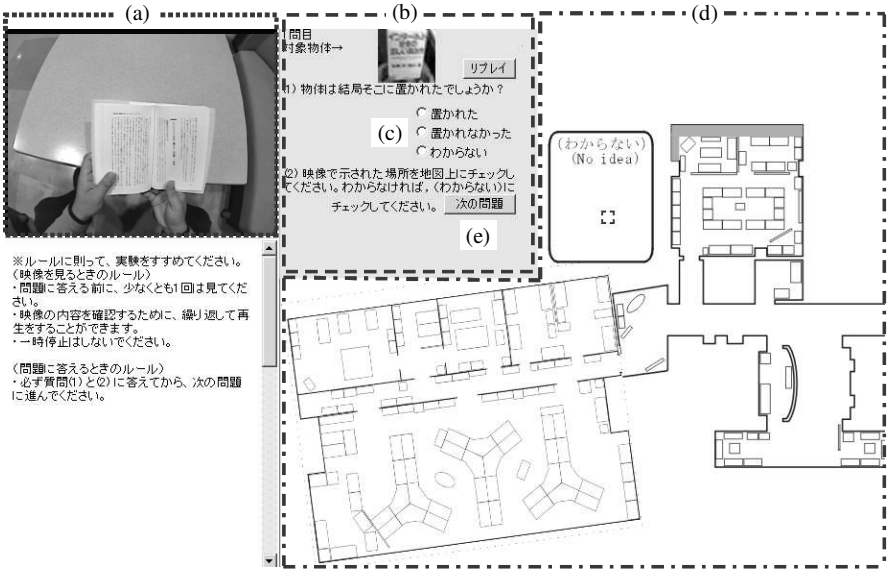

Figure 6: Voting Interface.
(a) Video display area, (b) Questionnaire display area,
(c) Action voting forms, (d) Place voting map,
(e) Questionnaire feeding button

Experimental videos can be classified into several kinds of groups with different rules. Videos displaying a common place with all levels of view size were classified into any of 10 place groups. Each of the 10 view size groups consists of 10 videos of a common view size level in 10 different places. In the case of focusing a lens used at the recording, each lens group was made up of videos extracted from a common source video recorded by the lens.

To make experimental videos simulating the results of the vision-based object recognition function employed in *I'm Here!*, the following segmentation rules were applied to trimmed source videos:

1) The end-point was determined when either of the following requirements was satisfied:

(a) The occluded area of the target object was over 30% of the entire object region. The occlusion was caused by the hand region holding the object or the frame of the trimmed video.
(b) The visible areas of the target object were under 2% of the entire view area.

2) The start-point of experimental videos was determined by applying the following rule separately to each lens group:
(a) The video trimmed to the smallest view size in each group was considered as a standard in the group. A common start-point of all videos in the group was determined to make the time length to the end-point of the standard video, preliminarily determined, should be 5 second.

Each experimental video was displayed on a browser. The voting interface illustrated in Figure 6 consists of video display area (Figure 6(a)), questionnaire display area (Figure 6(b)), action voting forms (Figure 6(c)), place voting map (Figure6(d)) and a questionnaire feeding button (Figure 6(e)). The interface enabled each subject both to replay a experimental video and to answer questionnaire.

The questionnaire was denoted as below:

1) How do you think that the target object was really placed there? Vote with following options:
a) It was really placed on there.
b) It was not placed on there.
c) I have no idea.

2) Where do you think that the target object was placed? Mark there on the map if you can think any proposed place, or you should mark on "no idea."

To answer the first question, each subject should decide whether the experimenter really placed or did not place the target object on any place displayed in the experimental video. If it was difficult for the subject, he could answer that he had no idea. A cognitive estimation at viewing a positive/negative action of placing the target object should be found through the question. Action recognition patterns of each subject are listed in Table 1.

		Actual action placing a target object	
		Positive	Negative
Answer by each subject	Positive	TRUE-POSITIVE	FALSE-POSITIVE
	Negative	FALSE-NEGATIVE	TRUE-NEGATIVE
	No idea	No idea (positive case)	No idea (negative case)

Table 1: Action recognition patterns.

The second question mentions cognitive estimations of each subject at viewing a place displayed in a viewpoint video. The map displays arrangements of fixed furniture in the laboratory environment, such as walls, columns, doors, bookshelves, desk and copiers. The subject allowed answering a number of proposed places up to 30. He also allowed answering "no idea" if the subject could not focused the proposed places.

When the subject marked those voting forms, he/she went to the next questionnaire. An experimental trial consists of a number of questionnaires displaying experimental videos belonging in a lens group. Questionnaires in a trial were ordered by view angle level of experimental videos into ascending sequence. A subject totally tried 3 trials with more than a day of interval between them. Each of first and second trials consists of 30 questionnaires, and the third trial consists of 40 questionnaires. Through all experiments for a subject, trials were ordered by viewing angles of lenses into ascending sequence so that all experimental videos were ordered into ascending sequence of view angle.

3.2 Results

From questionnaires written by subjects, results of action recognition and location recognition were counted on view angle basis. Answers to videos, which sources were taken in location 3 and 8, are not included for counted results because of some significant gap of contexts between sources with different lenses. Questionnaires about 80 videos consisting of 10 view angle conditions, generated from 4 sources of positive case and 4 of negative case, were totally counted.

Figure 7 illustrates subject number ratio of action recognition separated into TRUE-POSITIVE (Figure 7 (a)), FALSE-NEGATIVE (Figure 7 (b)), FALSE-POSITIVE (Figure 7 (d)) and TRUE-NEGATIVE (Figure 7 (e)). TRUE/FALSE means that a subject recognized action of placing a target object in each video correctly/incorrectly. POSITIVE/NEGATIVE means contents of the subject's answer whether the action was actually occurred/not occurred. We found that the ratios of TRUE-POSITIVE and TRUE-NEGATIVE increase and those of FALSE-POSITIVE and FALSE-NEGATIVE decrease with increasing view angle condition. When the actual action was positive, the ratio of "No Idea" becomes to decrease with

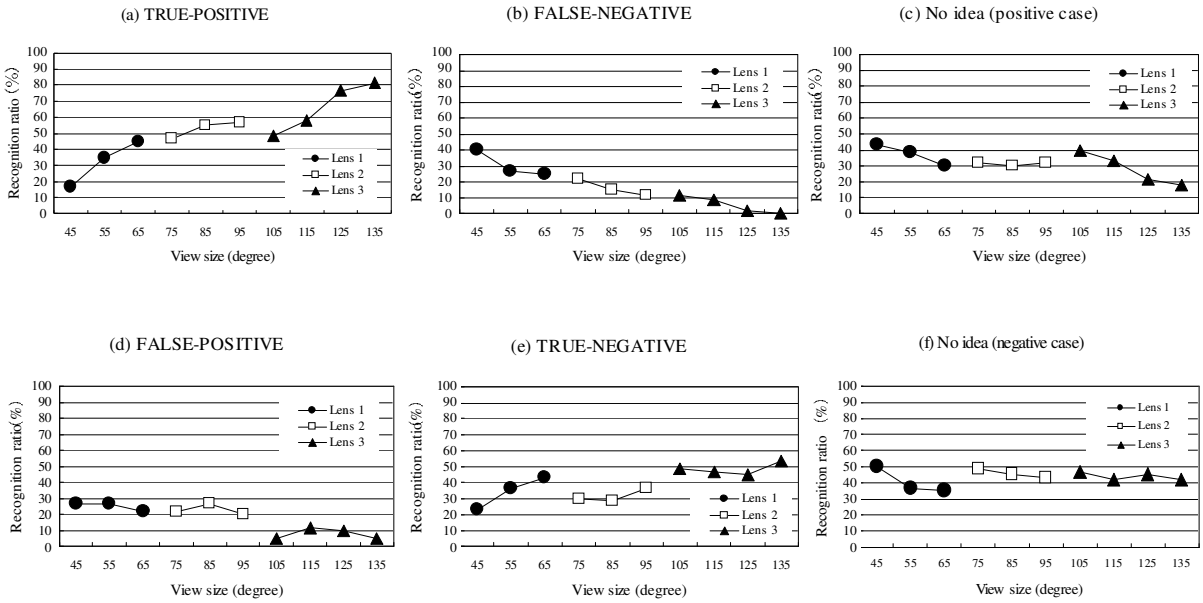

Figure 7: Results of action recognition

increasing the view angle condition as illustrated in Figure 7 (c). In the case that the actual action was negative, the ratio of "No Idea" illustrated in Figure 7 (f) almost remained unchanged.

Two evaluations of location recognition, tendency of narrowing off the number of candidate locations and accuracy of location recognition, were analyzed. The number of candidate locations includes all number of voted locations without the case that subjects answer that they have no idea.

From Figure 8, the number of candidate locations was well narrowed down by setting view angle condition to 115. Figure 9 (a) illustrates that the true results of location recognition were saturated from 125 of view angle condition. The false results illustrated in Figure 9 (b) almost decreased to 0 % when the view angle was set to 125. Figure 9 (c) shows that the results of "No idea" in location recognition decreased in small steps with increasing the view angle condition.

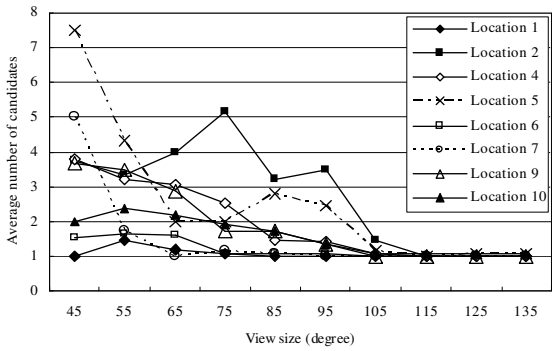

Figure 8: Narrowing down of candidate locations

4 Discussion

With experimental results, we define required/ sufficient view angle conditions to support a user's object finding task. When the view angle was 115, true graph of location recognition (Figure 9 (a)) was saturated with false graph, no idea graph and narrowing down graphs of location recognition (Figure 8, 9 (b) (c)) were asymptotically stable at each rock-bottom value. From those results, the required view angle condition should be set to 115. On the other hand, the sufficient view angle condition should be set to 125, because the true positive graph of action recognition was saturated in the setting.

There were, however, some videos with sufficient view angle that subjects could not recognize action placing the target object correctly. The action recognition failure of FALSE-POSITIVE pattern decreases reliability of *I'm Here!*, and that of false negative pattern decreases effectiveness of *I'm Here!*.

We found that subjects estimated the action placing the target object with its preparative action, even if they could not recognize the action

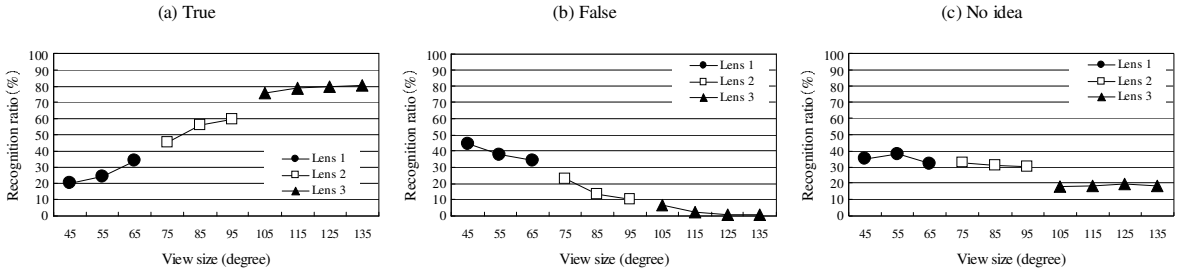

Figure 9: Results of location recognition

segmentation point. The gradual expansion of TRUE-POSITIVE graph (Figure 7 (a)) supports the presence of the estimation.

The action estimation, at the same time, causes the action recognition failure of FALSE-POSITIVE pattern. We named the negative effect of action estimation as "pseudo action segmentation." Actually, some subjects mistakenly recognized the video of negative action displaying confusing movement of hand with the target object as a positive action. It is also difficult for the vision-based system to distinguish accurately the case of action segmentation from that of pseudo action segmentation unless the action segmentation point can be certainly in sight of the camera. To constrain the pseudo action segmentation, *I'm Here!* should display any additional viewpoint video displaying other kind of action, such as "going to stand up" or "starting walking" before placing the target object. The user will be able to estimate the action "not placing the target object" by those contexts.

5 Concluding remarks

We have proposed a method to support a user's object-finding task using a first-person video of the user recorded when he/she last held a target object. If the user watches the video, he/she can vividly recall the location where he/she last placed the target object.

In this paper, we investigated the effective view angle of a lens for the first-person video. Through experiments with subjects, we found that a lens with a range of view angle from 115 to 125 (degrees) is better suited for recording the first-person video. On the other hand, some experimental results showed that subjects misunderstood the placing action illustrated by first-person videos. For example, subjects saw a target object placed in a location, but in actuality, no target object was placed in that location. This case, called a "FALSE-POSITIVE" is caused by a pseudo action segmentation illustrated in the video which looks like a placing action. Such false-positives can be avoided if a user views the video with an additional first-person video illustrating a post-event such as 'standing up' or 'walking'. In this case, the user will easily understand that the target first-person video illustrates no placing action.

We plan to design and evaluate methods to identify those post-events. We also plan to implement our wearable augmented memory system to support a user's object-finding task with an effective view angle lens.

References

[1] L. Davenport. Order from Chaos. Three Rivers Press, New York, 2001.

[2] M. Lamming and M. Flynn. Forget-me-not: Intimate Computing in Support of Human Memory. *Proc. FRIENDS21 : International Symposium on Next Generation Human Interface*, 125-128, 1994.

[3] R. J. Rhodes. The Wearable Remembrance Agent: a System for Augmented Memory. *Proc. International Symposium on Wearable Computers (ISWC'97)*, 123-128, 1997.

[4] T. Ueoka, T. Kawamura, Y. Kono and M. Kidode. *I'm Here!*: a Wearable Object Remembrance Support System. *Proc. Fifth International Symposium on Human Computer Interaction with Mobile Devices and Services*, pp.422-427, 2003.

[5] T. Ueoka, T. Kawamura, Y. Kono and M. Kidode. Functional Evaluation of a Vision-based Object Remembrance Support System. *Proc. IEEE International Conference on Multimedia and Expo (ICME'2004)*, 2004.

[6] M. Shinnishi, S. Iga, F. Higuchi and M. Yasumura. *Hide and Seek* : Physical Real Artifacts which Responds to the User, *Proc. World Multiconference on Systemics*, Cybernetics and Informatics (SCI'99/ISAS'99) Vol.4, pp.84-88, 1999.

[7] Y. Ikei, Y. Hirose, K. Hirota and M. Hirose. "*iFlashBack*": A Wearable System for Reinforcing Memorization Using Interaction Records. Human-Centred Computing, Vol.3, pp. 754-758, 2003.

[8] D. A. Newtson. Foundations of attribution: The perception of ongoing behavior. J. Harvey, W. Ickes, & R. Kiss (Eds.), New direction in attribution research (vol. 1). Hillsdale, NJ: Lawrence Erlbaum Associates, 1976.