NAIST-IS-DT0261008

Doctor's Thesis

I'm Here! : a Wearable Memory Augmentation System to Support an Object-finding Task

Takahiro Ueoka

March 18, 2005

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctor's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Takahiro Ueoka

Thesis Committee: Masatsugu Kidode, Professor (Supervisor)
Yasuyuki Kono, Associate Professor
Kunihiro Chihara, Professor
Yasuyuki Sumi, Associate Professor

(Revised in July 24, 2006.)

I'm Here! : a Wearable Memory Augmentation System to Support an Object-finding Task*

Takahiro Ueoka

Abstract

This thesis discusses a wearable memory augmentation system used to make a person's object-finding tasks in his/her everyday life efficient. A person tends to have to accomplish a task to find a target object held and moved by him/her in an everyday environment. The task, named an object-finding task, is caused by ambiguity of human memory activity such that a person forgets where he/she last placed a target object. If the person recalls where the target object has been placed last, he/she can effectively find the object without wasting time.

To support a user's object-finding task, this thesis proposes a wearable memory augmentation system "*I'm Here!*" which displays a video of his/her viewpoint recorded when the target object was last held by him/her. *I'm Here!* employs a vision-based object recognition method to identify the object observed in viewpoint images captured by a head-mounted camera device. *I'm Here!* has a simple system construction without any additional device exclusive for object identification. Preliminarily, the user wearing *I'm Here!* only has to register appearances and names of target objects. The user then can retrieve the video of a target object by only selecting the name of the object.

This thesis discusses the following two topics to evaluate the system design of *I'm Here!* : (1) The contribution of *I'm Here!* in supporting object-finding tasks through experiments with subjects using first-person videos, (2) The implementability of *I'm Here!* through functional evaluations of novel camera devices which play an important role in recording the first-person video.

* Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT0261008, March 18, 2005.

(1) To evaluate the significance of the *I'm Here!* system design, subjects were given tasks to find target objects in their everyday environment using the experimental *I'm Here!* system. The experimental *I'm Here!* system includes a mounted video database preliminarily constructed with a controlled object-recognition ratio. On the other hand, to evaluate the reliability of first-person videos recorded and retrieved by *I'm Here!*, subjects watched experimental first-person videos and answered questions about how each video might be useful in finding the target object. Each experimental first-person video was preliminarily created by the method that *I'm Here!* employs under the condition of a view size parameter. Through these experiments, parametric requirements were included in the object recognition ratio and the view size of the first-person video.

(2) I have developed novel camera devices named the "*ObjectCam*" series, needed for the object recognition function. These devices are designed to extract images of a target object from first-person videos by eliminating background regions in real time. Two experiments were conducted using the *ObjectCam* series. In one experiment, a histogram-based object recognition method with a prototype *ObjectCam* was implemented and demonstrated. As a result, a sufficient object recognition ratio to support object-finding tasks was marked. In the other experiment, *ObjectCam2*, the latest version of my cameras, was tested to see if it could extract target object images accurately under an environmental condition in which actual object-finding tasks can occur. As a result, I found that *ObjectCam2* possesses the ability to extract target object images under a practical environmental condition.

With these results, this thesis shows that the system design of *I'm Here!* is effective and practical when used to support a user's object-finding task in an everyday environment.

Keywords:

wearable computing, augmented memory, object-finding task, object recognition, camera device

I'm Here!：物探しを支援するウェアラブル拡張記憶システム*

上岡 隆宏

内容梗概

本研究の目的は，物探しタスクを効率化するウェアラブル拡張記憶システムを設計・評価することである．物探しタスクとは，人が物体を見つけるために日常生活環境の中を探しまわるタスクであり，その対象は人が持ち運んで使用する把持物体である．物探しタスクが発生する原因は，自分自身が把持物体を最後に置いた場所を思い出せなくなることにある．物探しタスクのために浪費する時間が短縮されれば，人は余った時間を他の有意義なタスクに割り当て，日常生活をより豊かにすることができる．

本研究で提案するウェアラブルシステム*I'm Here!*は，ユーザの視野映像を蓄積・検索するウェアラブル拡張記憶システムを応用し，対象物を最後に置いた時点の映像をユーザに提示することで物体を置いた場所の想起を支援する．*I'm Here!*は，頭部装着型カメラデバイスで撮影した視野映像を利用したビジョンベースの物体認識を行い，物体認識の結果に基づいて対象物体を最後に置いた時点の映像を記録する．そのため，*I'm Here!*は物体認識専用デバイスを必要とせず，単純なデバイス構成で実現される．*I'm Here!*の導入に際して，ユーザは物体や環境に認識用のタグを添付するなどの大きなコストを払う必要はなく，ウェアラブルデバイスを身に付けるだけでよい．あらかじめ簡単な操作で把持物体の名前と画像特徴を登録しておくことで，ユーザは名前の選択をするだけで把持物体が最後に置かれた時点の視野映像を見ることができる．ユーザは，提示された視野映像に含まれるコンテキストを理解することで，ユーザ自身が最後に対象物を置いたイベントに関する記憶を補強し，対象物が置かれている場所を見当付け，物探しを効

* 奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DT0261008，2005年3月18日．

率的に行うことができる．本研究では，(1) *I'm Here!*のシステムデザインを評価するために，被験者による仮想的な物探し実験を通じて*I'm Here!*が物探しタスクを効率化することを検証した．また，(2) 現実的なシステム構築の可能性を検証するために，*I'm Here!*が視野映像をインデクシングするために重要な役割を担うカメラデバイスを実装し，評価を行った．

(1)　*I'm Here!*によって構築される視野映像データベースを仮想的に準備し，被験者実験を行った．*I'm Here!*のシステムデザインがもたらす物探し支援効果を評価するために，映像構築時のパラメータとして (a) 物探しタスクを効率化するために必要な物体認識率，(b) 物探しを支援する提示映像の視野角条件に着目した．(a) では，被験者による仮想的な物探し実験を通じて，*I'm Here!*が物探しを効率化することが示された．同時に，*I'm Here!*が用いる物体認識手法が満たすべき物体認識率の要件が導出された．(b) では，仮想的に作成した視野映像を被験者がアンケートによって評価する実験を行った．その実験によって，被験者が限定的な視野角条件が課せられた視野映像によって被験者が物探しに必要なコンテキストを認識するかどうかを評価した．結果として，視野映像がユーザの物探し支援に高い効果を発揮するために必要な視野角条件が定量的に導出された．

(2)　本研究では，物探しタスクが発生する環境において把持物体を実時間で認識するために，視野映像から把持物体の画像を背景と分離し抽出する機能を持つ頭部装着型カメラデバイス ObjectCam シリーズを開発した．ObjectCam シリーズを利用して，(a) 物探しを支援するために必要な物体認識機能の実装・評価，(b) *I'm Here!*が使用される環境条件においてロバストに把持物体の画像を抽出する機能の実装・評価が行われた．(a) に対しては，プロトタイプ ObjectCam を利用し，画像ヒストグラムに基づいた物体認識手法を実装した．日常生活で用いられる把持物体を認識する実験を通して，物体認識手法が物探しを支援する動作精度を持つことを確認した．(b) に対しては，改良型の ObjectCam2 を利用し，物探しが行われる日常生活の環境条件下において物体画像の抽出精度を評価する実験を行った．その結果，ObjectCam2 が実践的な環境条件下において物体画像の抽出能力を持つことが確認された．

これらの結果，日常生活環境においてユーザの物探しタスクを効率化するウェアラブルシステム *I'm Here!*が構築可能であることがわかった．また，I'm Here! が物探しを効果的に支援するための物体認識率・映像視野角に関する要件が求め

られた.

キーワード

ウェアラブルコンピューティング，拡張記憶，物探しタスク，物体認識，カメラデバイス

Acknowledgements

I have been supported by a number of people through my research activities. I want to take this opportunity to thank them.

I would like to first thank you Prof. Masatsugu Kidode for his encouraging and accurate advices. I would like to thank Prof. Kunihiro Chihara. I would like to thank Associate Prof. Yasuyuki Sumi. I would like to thank Associate Prof. Yasuyuki Kono for his detailed support. I would like to thank Mr. Tatsuyuki Kawamura for giving me a lot of stimulation and encouragement. I would like to thank Mr. Shinichi Yoshimura and Mr. Shigeyuki Baba for developing the nobel camera device with me as a collaborative research.

Next I would like to thank all members in Artificial Intelligence Laboratory of NAIST. With their dedicated collaboration, I can get worthwhile experimental results.

This research is supported by Core Research for Evolutional Science and Technology (CREST) Program "Advanced Media Technology for Everyday Living" of Japan Science and Technology Agency (JST).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Goal of Thesis

Currently, information services provided by computers are utilized on an everyday basis to perform tasks more efficiently and with less effort. Computers process vast amounts of data to achieve information services such as the creation and storage of documents and the retrieval of documents, music and videos. Nonetheless, some , people still spend considerable time looking for objects (in object-finding tasks) which are neither exciting nor meaningful. People often try to find a target object such as a key, or some other target object that they misplaced and cannot find, but need immediately. Finding a target object often wastes a great deal of time in the daily life of many people. Information technologies should be used to support such troublesome tasks so that human life will be easier and more comfortable.

The main goal of this thesis is to propose and evaluate a design for a wearable computing system supporting a user's target object-finding task by augmenting human memory activity. A number of related studies have challenged to support a user's object-finging tasks in his/her everyday life. Winograd and Soloway (1986) investigated relationships between objects and locations from the perspective of human memory. Shingaki et al., (2003) investigated how human loses objects in

everyday environment with focusing cognitive process. Terai and Miwa (2004) analyzed human's cognitive process of object-finding task with focusing similarity between the process of object-finding task and the process of insight problem solving.

This thesis mentions evaluations of a system design for supporting a user's object-finding task with focusing two fundamental points. The first point requires an evaluation of an object recognition ratio to create an effective memory-aid. The other point requires the evaluation of a view size for a first-person video, in other words, a video captured from the view point of a user.

1.2 Augmented Memory

This section discusses Augmented Memory by briefly reviewing the literature pertaining to studies of wearable computing and Augmented Memory. Needing to find an object is caused by failures of human memory activities. Psychological studies have shown that the human brain tends to cause mistakes in the process of memory activities (Nickerson, R. S. and Adams, M. J. 1979). Human memory consists of two basic parts, a short- term memory and a long- term memory. Long-term memory includes memory activities with episodic memories. Tulving, E. and Thomson, D.M. (1973) have noted that episodic memory contains a person's experiential episode, which differs from semantic memory, such as the meaning of words. Failures in memory activities sometimes last long enough so that a person many never be able to recall a particular episodic memory required for object finding tasks. The required episodes can be recovered only if they have been recorded on a certain media or if they have been shared with other persons.

Augmented Memory, a concept used to support augmentation of a person's memory activities, has been proposed in recent years (Rhodes, B. 1997). An Augmented Memory system has been studied in area of wearable and ubiquitous computing. Wearable sensors can be used continuously to accumulate intimate information included in videos of the user's viewpoint. For example, an Augmented Memory system with a head-mounted camera accumulates video of a user's viewpoint into a wearable computer as his/her visual experiential memory.

In the case that the user wants to recall his/her lost memory, he/she can utilize the Augmented Memory system to replay a video of his/her viewpoint required for the recollection. The possibility of implementing the concept of Augmented Memory has increased as recent technology has progressed. If a person records a video from his/her viewpoint into storage medias during 16 hours in a day with encoding in 1Mbps, the size of the accumulated video data will be approximately 190 TiBs (Terabinary Bytes) after 80 years later. Storage media will become cheaper and smaller as PCs become cheaper and smaller. A personal storage media with 190 TiBs of memory will be reasonably priced in the years ahead.

1.3 Wearable and Ubiquitous Computing for Augmented Memory

Several studies discuss achieving memory-aid applications by using wearable computing environments. Mann, S. (1997) utilized sensors and cameras to record everyday life. In studies by Kawashima, T. et al., (2002) and Toda, M. et al., (2003), a wearable video segmentation system that recognizes location and human action has been studied. Jebara, T. et al., (1998) and Shiele, B. et al., (1999) proposed *DyPERS*, a system to record andreplay a user's visual and auditory scenes. *VizWear-Active*, a shoulder-mounted camera system, supports a user's ability to recall events such as recognizing andtracking human faces (Kato, T. et al. 2002). Kidode, M. (2002) produced a project named *WIPS* to integrate wearable media technology. Actually, my studies mentioned in this thesis has been joined in *WIPS*.

Memory-aid systems based on ubiquitous computing technologies also have been studied. The concept of ubiquitous computing was first proposed by Weiser, M. (1991). Lamming, M. and Flynn, M. (1994) proposed a PDA-based portable system to aid episodic memory utilizing ubiquitous sensors in a laboratory. Sumi, Y. et al., (2002) proposed a method using wearable and ubiquitous sensors to generate video-based experience summaries. This method is based on the notion of an interaction corpus, which illustrates human behaviors andinteractions among

humans and artifacts.

In parallel with application studies, evaluations of the effectiveness of Augmented Memory have been the focus in some studies. Ueoka, R. et al., (2001) discussed an applicable view angle for constructing a video-based experience database. Narita, N. et al., (2000) have been investigated a effectiveness of wide-screen display by comparing sevral aspect ratios. *MyLifeBits* system utilizes a neck-worn camera named *SenseCam* to store pictures illustrating a user's lifetime (Gemmel, J. et al., 2004). Lenses for *SenseCam* in different width of view angle has been compared so that recorded pictures can illustrate what the user is seeing. This thesis focuses a proper view angle of lens to capture first-person videos for supporting a user's object-finding task.

1.4 Overview of Thesis

In this thesis I propose a system named *I'm Here!*, as a wearable system to support an object-finding task using Augmented Memory. This thesis contributes to create a wearable system that may be utilized by a user with comfort in his/her everyday life.

I'm Here! has been developed as a module of a *SARA* framework which utilizes Augmented Memory to organize a user's everyday memories just like using a photo album to recall memories (Kono, Y. et al., 2003). *SARA* contains not only *I'm Here!*, but also the following systems. Ubiquitous Memories provides a video accumulate andretrieval interface based on real-world objects with attached RFID tags (Kawamura, T. et al.,2003a). Residual Memory is a vision-based video retrieval system for recalling past events at certain locations (Kawamura, T. et al., 2001). *Nice2CU* supports a user's recall events associated with acquaintances (Kawamura, T. et al., 2003b). The study of the *SARA* framework discusses that these modules should be integrated complementarily to allow users to access augmented memories freely and alternately.

Figure 1.1 denotes the logic structure of this thesis. Chapter 2 discusses the concept of the proposing a wearable system by comparing related works. In Chapter 3 discusses the overview of the proposed system design and details

subject-centered experiments to establish the objective parameter values needed in an ideal system. Chapter 4 discusses in detail the implementation related to the most important modules of the proposed system. Experimental evaluations of the implemented modules are also explained in chapter 4. Chapter 5 discusses contributions created by this study, and concludes this research with discussions about extended functions to make the wearable Augmented Memory system more useful in the future.

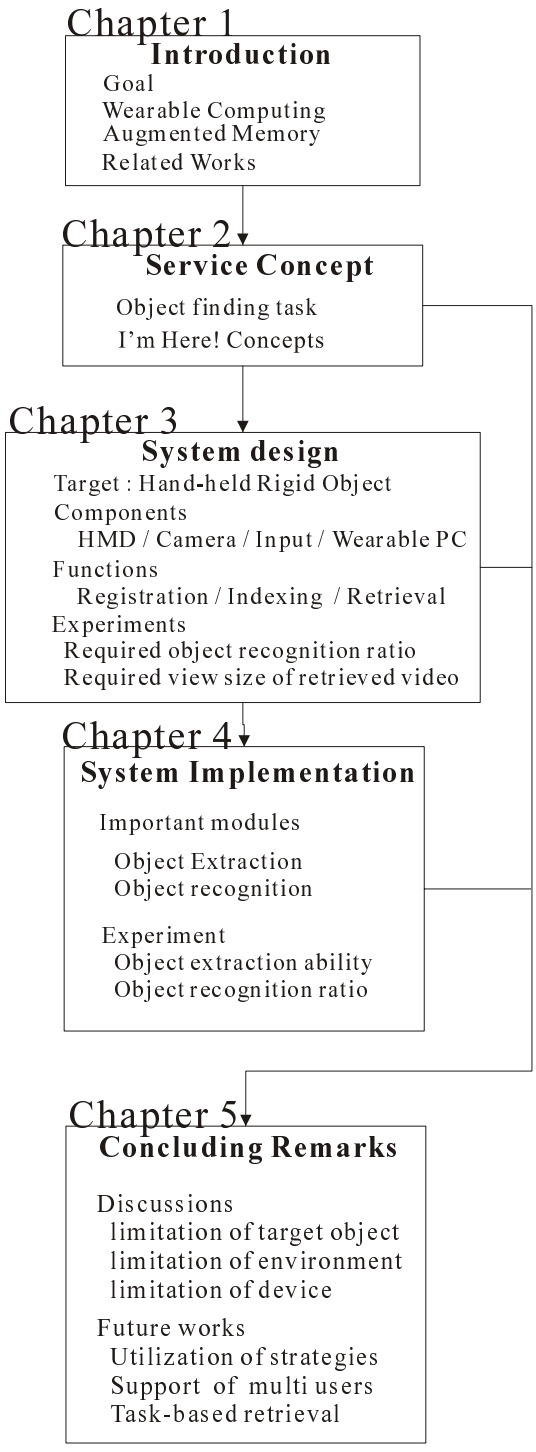

Figure 1.1. Overview of thesis

Chapter 2

Augmented Memory to support object-finding task

This chapter mentions a concept design of wearable system named *I'm Here!*. *I'm Here!* is designed just for supporting memory activities triggered by objects on the basis of the concept of Augmented Memory previously denoted in Chapter 1.

2.1 Problems of Object-Triggered Memory Activities

Many objects trigger memories in a user's everyday life. Shingaki, N., et al. (2003) note that the number of object a person owns in his/her everyday environment is larger than 10,000. In an empirical case of a student, the number of hand-held rigid objects owned by him is about 180. The following presents example problems of memory recollections triggered by objects.

[A person (e.g. "Tom") can't recall...]

1. a nostalgic episode or a memorandum related to an object he handles
2. a person to whom he has lent an object he handles
3. a person from whom he has borrowed an object he handles

4. how he has used an object he handles last time

5. what objects he has to take with himself for a trip

6. what objects he needs to do a certain task

7. where he placed an object

8. the price of an object sold in another shop to compare it with another similar one sold in front of him

In this study, I notate the case 7 denoted previously, a failure of recalling location where a person placed a target object in his/her everyday life. There are many tasks each of which needs a specific object in everyday life. If a person cannot recall where the specific object is placed, he/she comes to search for the object before solving the task. In this paper the task searching for object is named as an object-finding task. People spend their resources such as time and energy to object-finding tasks in their everyday life. Actually, Davenport, L. (2001) notes that an average businessperson spends 150 hours in one year for the object-finding tasks. The object-finding task can also occur before another memory activities denoted in previous case 1, 2, 3, 4 and 8. People would be able to finish the object-finding task immediately by supporting recollection where the object was placed.

2.2 object-finding Support with Augmented Memory

Figure 2.1 illustrates a conceptual image of *I'm Here!* supporting object-finding task. *I'm Here!* diplays appropriate video to the user when it is required for the user's memory recollection. For example, *I'm Here!* records a first-person video when he/she places a pen case onto a table automatically. *I'm Here!* also annotates a name of the pen case to the video displaying the pen case held by the user. The name of the pen case is preliminarily registered by the user. The user can watch the last video only by feeding a name of the target pen case to I'm Here. Even if the user forgets where he/she placed the pen case, he/she can recall the place by watching the video.

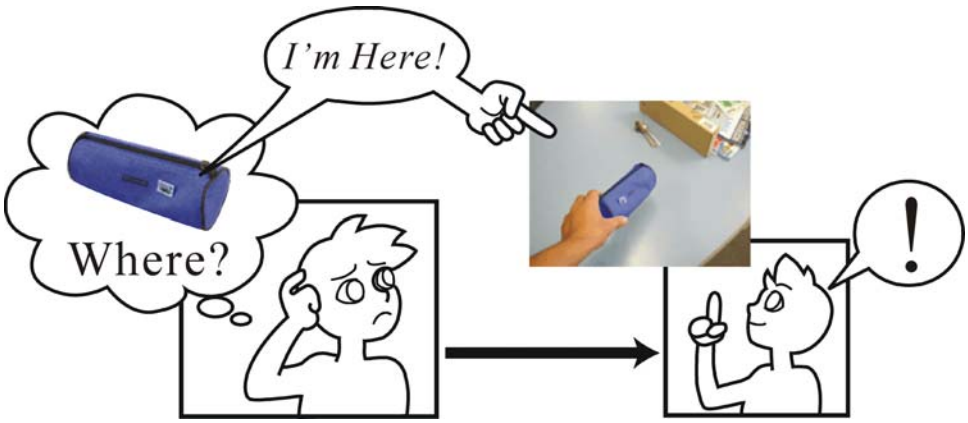

Figure 2.1. Concept image of *I'm Here!*

I'm Here! realizes an Augmented Memory with Experience Database and Object Database. The Experience Database coinsists of videos of a user's viewpoint with an index of objects. The Object Database accumulates the name and visual features of each registered object for vision-based object recognition method. During the user performs several tasks using registered objects in his/her everyday life, *I'm Here!* continuously stores his/her experiences triggered by objects into the Experience Database. The scene that the user places an object on a certain place is also included in the Experience Database. To index the Experience Database with names of objects, *I'm Here!* recognizes the object observed in the Experience Database on the basis of vision-based recognition method with the Object Database.

It commonly happens that a person in execution of a task comes to have to do the other task preferentially. In this paper I call the situation that a task is interrupted by the other task as "task interruption." The object-finding task tends to occur in association with the task interruption. Following story illustrates an example of the situation :

> When a student named **A** had a few break with a cup of coffee, he was called by a research assistant for supporting an urgent experiment. **A** placed his cup on a window ledge near him and moved to the experimental room to support the experiment immediately. After that, **A** performed other tasks in a row without caring about his coffee cup placed by himself. When **A** was suddenly tempted to drink a coffee, he found himself forgetting where the cup was placed. **A** searched

for the cup in everywhere he might have placed the cup, and finally found the cup on the window ledge with spending more time and energy.

In the example illustrated above, the task interruption caused **A**'s lack of attention to the cup. The failure in memorization of the location where the cup was placed has been triggered by the lack of attention.

I'm Here!, mentioned above, supports such failures in a user's memory activities. During a user wears *I'm Here!* system, *I'm Here!* continuously captures video of the user's viewpoint into Experience Database with index of observed objects which are preliminarily registered in Object Database. The function of capturing and indexing aids the user's lack of attention to the target object, and also aids the user's failure in memorization. The failure in retention and recollection, the other memory activities, are also aided with the Experience Database constructed on a nonvolatile storage of *I'm Here!*.

2.3 Related Works

Shinnishi et al. (1999) have proposed "Hide and Seek," an object-finding support system based on a concept that the target object responds by sound to a user's call. An active ID tag used in Hide and Seek consists of an infrared signal receiver and a speaker. Each active ID tag is attached to each target object. A Controller of Hide and Seek has an infrared signal transmitter and a microphone. When the user pronounces the name of the target object, the controller transmits infrared signal specified to the ID of the object. In respond to the signal, The active ID tag makes a sound. The user can understand the location of the target object from hearing the controlled frequency of the sound depending on physical relationship between the controller and the tag. Hide and Seek is expensive to introduce and maintenance of the system in environmental improvement, such as attachment of tags to target objects and charging tag batteries. *I'm Here!* can be introduced and maintained less expensive than Hide and Seek. Because *I'm Here!* only requires operations for registration of target objects to a user in the system introduction.

Ikei et al. (2003) proposed "iFlashBack," the wearable video retrieval system to support a user's ability of memorization. iFlashBack consists of a head-mounted camera, a head-mounted display and RFID tag/tag reader. iFlashBack automatically captures

video of the user's viewpoint and segments the scene of the user's handling a target object with an RFID tag attachment. When the user releases the target object, iFlash-Back displays the last segmented scene again to the user. The user's memorization is reinforced by a rehearsal effect caused by reviewing the scene. iFlashBack assumes that the target objects have been preliminarily attached to target objects. *I'm Here!* is more practical than iFlashBack with direct displaying of the don't use any tags attached to objects.

Chapter 3

I'm Here! system design

This chapter mentions a system design of *I'm Here!*. The system design is for single user only. *I'm Here!* helps him/her finding a hand-held rigid object placed by himself/herself in his/her everyday environment.

3.1 Overview of *I'm Here!* System Design

Figure 3.1 illustrates a design for the object-triggered memory augmentation system. There are three phases in using the system : registration, indexing, and retrieval. The example system is composed of a PC, a Head-Mounted Display (HMD), a wearable camera that can capture a first-person video, and an input device/interface. In the example, the system records a video as experience data by using the wearable camera. The wearable camera is also employed to detect an event such as when a trigger object is handled, and to identify the object. The system can get an identification (ID) number after the object identification process.

Registration Phase

- **State A:** The user registers a trigger object as preparation for recollection of his/her experiences.
- **State B:** The user enters the data, e.g. an object name, an owner name, and attributes like a category class by using the input device/interface.

Indexing Phase

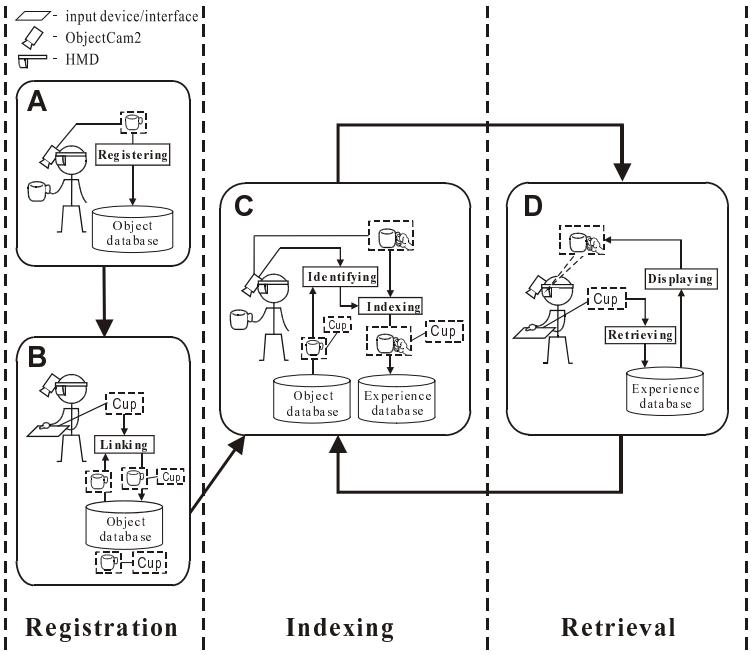

Figure 3.1. Phases of an object-triggered memory augmentation system

- **State C:** The system continuously records a video by using the wearable camera. When the user handles a certain registered object, the system detects the event and identifies the object by using the input device/interface. The system then indexes the ID number of the object to the experience data.

Retrieval Phase

- **State D:** The user enters a query, e.g. the name of a registered object, by using the input device/interface. The system then retrieves a video associated with the query. Finally, the system displays the video to the HMD. The user can use this type of retrieval even if he doesn't have the registered object at hand.

I'm Here! handles a user's first-person video in his/her everyday life to construct Augmented Memories for supporting his/her object-finding task. The user has to wear a wearable system of *I'm Here!* continuously. *I'm Here!* is designed on the basis of following requirements for being used in the user's everyday life.

- Less difficulty in operating *I'm Here!* interface

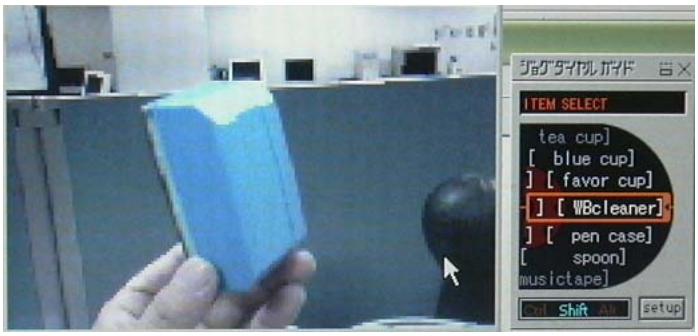

Figure 3.2. An image shot of memory retrieval with *I'm Here!*

In Phase A on Figure 3.1, the user holds a target object and gazes it from several angles with rotational operations. Because these rotational operations are freed from the constraints of rotational direction and operational order, the user can instantly carry out the phase although he/she first wears *I'm Here!*.

Phase B and D on Figure 3.1 make the user to manipulate a menu list displayed in the head-mounted display (Figure 3.2) using an input device. The manipulation consists of simple motions, shifting menu to sight on a target item and selecting the target item. It is easy for peoples who are familiar with operations of several electric appliances to become friendly with the simple manipulation.

- Less behavioural obstruction in a user's everyday life

  I'm Here! constructs Augmented Memories automatically. Phase C in Figure 3.1 constitutes a larger part of the user's everyday life than phase A, B, and D. In phase C, the user has only to do his/her everyday activities with wearing the wearable system. The user is no cause for being conscious of the system's behaviours, capturing a first-person video and constructing Augmented Memories. By adopting the system design that I explain in this section, the major factor of *I'm Here!* that obstructs the user's everyday activities has been only size and weight of the wearable system.

The system design of *I'm Here!* is appropriate to support object-finding task in a user's everyday life. On the other hand, it cannot be assured that *I'm Here!* constructs complete Augmented Memories for object-finding support. The completeness of Augmented Memories means that it certainly includes every scenes of each target object

lastly being held by the user. Following factors cause the incompleteness of Augmented Memories in *I'm Here!* :

- Vision-based method in object recognition

  I'm Here! employs a vision-based method in recognizing target object with its appearances displayed in a first-person video. It is difficult to achieve perfect accuracy in the vision-based method under everyday environments. The difficulty is caused by following factors :

  - Partial occlusions of an object image by the user's hand holding the object
  - Chromatic change of an object image relevant to illumination conditions in everyday environment

- Limited view area of head-mounted camera

  Augmented Memories with a first-person video bring merits which are mentioned in Chapter 2. However, a head-mounted camera device with limited view area cannot perfectly capture an entire scene of user's holding a target object from the user's viewpoint. Because of impracticability of observing the object, some partial scenes that the user holds the object off the view area would be dropped out from Augmented Memories. This problem is relevant to a user's perception of a video retrieved from Augmented Memories. The user gets to decide how the video is believable as a correct scene that he/she lastly holds the target object, even though Augmented Memories are imperfect.

In following sections I mention two sets of experiments. These experiments aim to investigate effectiveness of *I'm Here!* for object-finding support in relation to two parameters. One is an object recognition rate of the identifying module in Phase C on Figure 3.1, and the other is a view angle of head-mounted camera. I try to set desired value of each parameter using results of the experiments.

3.2 Effectiveness of *I'm Here!*

I have conducted experiments to evaluate the effectiveness of *I'm Here!* supporting a user's object-finding tasks. I have also evaluated both the positive and negative effects of the system on the behavior of the user by conditioning the accuracy of object

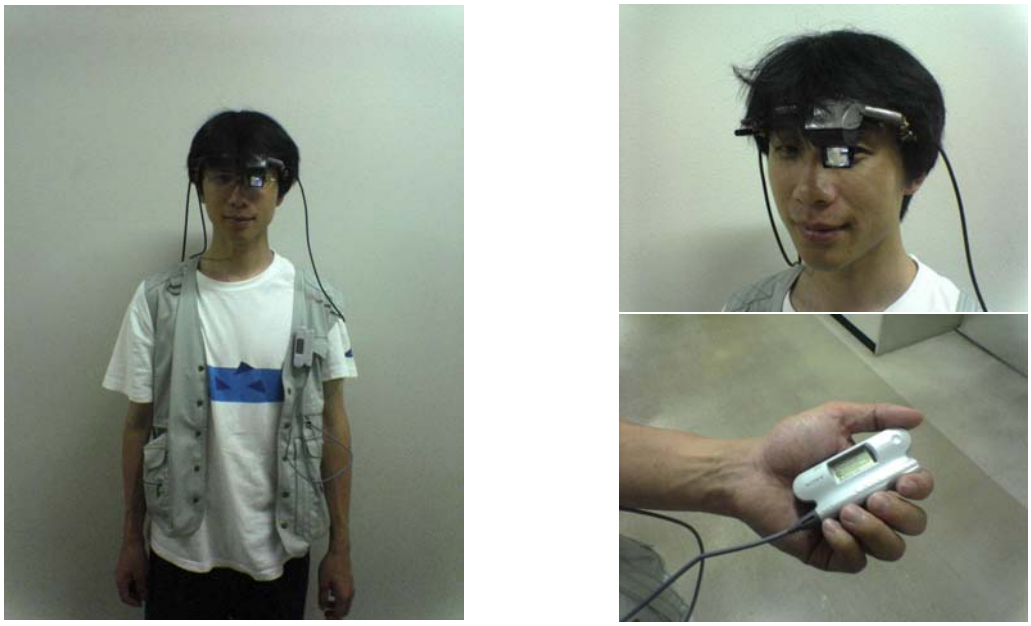

Figure 3.3. Hardware for the experiments

recognition. To survey the system's effects and contributions, I have configured the system and environment for the experiments by virtually setting the accuracy of object recognition.

3.2.1 Experimental Evaluation

Conditions

The wearable system for the experiments consists of a head-mounted display with CCD camera, a wearable PC, and a JogDial interface (Figure 3.3). Additionally, a pedometer is loaded to count the total number of steps in an experiment of the wearer. The experimental system simulates the retrieval phase. The CCD camera captures the first-person video only for displaying the real-time video to the user.

The Augmented Memories which were recorded in advance are loaded into the experimental system. The Augmented Memories are configured by applying a virtual object-recognition rate which is set as one of the following percentages : $\{0\%, 33\%, 66\%, 100\%\}$. The percentages represent the assortment of the positive and negative cases included in each trial of an object-finding task. The positive case represents the case where the

retrieved video displays the view of the virtual subject, the experimenter, and placing the target object in its actual place in the target-object placing task. The negative case represents the case where the video displays a view of the virtual subject placing an object somewhere other than where the object was originally and actually placed. The case where an experimental setting doesn't allow a subject to use the interface of the system is named as the without-system case.

The experiments were performed in my laboratory environment. Twenty-one preset places where the objects could possibly be placed were defined in advance. Twenty-one objects were prepared and each object was placed at a predetermined specific place. Two patterns of correspondence between the objects and the places are defined. These patterns also determine the sequence of placing the objects.

The subjects are seven male students of my laboratory who are familiar with the environment. Each subject carried out two experimental trials with the interval of a day between trials. The first and second trials for each subject have different patterns of object-place correspondence, and are common to all subjects.

Methods

A trial consists of 1) an object-placing task with a pattern of object-place correspondence, 2) memory-stressing tasks, and 3) an object-finding task. At the end of the second trial, each subject performed 4) an auxiliary data-extracting task. After a month of doing task 4, each subject is 5) interviewed.

• Task 1: The object-placing task

In this task, a subject wearing the virtual online version of the *I'm Here!* system places all twenty-one objects in a designated target area one by one via his shuttling between the starting point and the target place. At the starting point, the subject gets the object with information displayed on a card and a map. The card indicates the name of the object with its image, and the map indicates the location of the target place to put the object. A view image of the place for detailed annotation of the target place is attached to the map. The pedometer counts the total number of steps walked by the subject. Both the steps and the time required by the task are recorded.

• Task 2: The memory-stressing task

When the subject finishes task 1, he then performs the memory-stressing tasks before the finding task. The subject memorizes the location of twenty-five numbers

on a 5×5 array within ten seconds. Then he sequentially clicks grids of the array in ascending order of the numbers without displaying the numbers. Throughout this task each subject adds stress to his memory.

• **Task 3: The object-finding task**

In this task, the subject seeks three objects one by one while walking around the experimental environment. Before he starts walking, the card displaying the name and an image of the target object is shown to the subject. Next, the subject is allowed to operate the interface of the system only once, i.e., he selects an object with the Jog Dial interface, if the setting of the trial is not the without-system case. The virtual online system then displays the video previously associated with the object on the head-mounted display. Positive and negative cases in a task have been controlled as the setting of the task as below :

> *In the 1st experimental trial of subject P, the object-recognition rate is virtually defined as 33%. In P's object-finding task, the retrieved videos consist of one positive case and two negative cases, i.e., in a positive case the video shows the scene of placing the target object where it has actually been placed, and in each two negative cases, the video shows the scene of placing the target object where it had not actually been placed.*

The number of steps and the time required by the subject's walk are recorded in the same way that the object-placing task was done.

• **Task 4: The auxiliary data-extracting task**

In this task, the subject performs a short and long shuttle walk. Each walk consists of a starting, turning, and stopping motion, the same as the walk in the object-placing and finding tasks. The total steps and the time required by the walk are recorded as the normal walking data for the subject.

• **Task 5: The interview task**

In this task each subject is interviewed by an experimenter about his internal states of mind during the experiments. A subject answers some questions interactively, and the questions mainly consists of A) How he memorized the place of objects in task 1, B) What he was reminded of at the time the card indicating the name and an image of the target object is displayed in task 3, C) How he felt about the reliability of the displayed video in the case of using the system, and D) How he decided on his strategy to seek the target object.

Parameters

The parameters for calculating the quantitative contribution of the *I'm Here!* system are extracted from the recorded data. To extract the parameters, I assume the subject's walking model is denoted by the following variables : 1) W is the total number of steps, 2) L is the length of the optimum route to reach the given object, and 3) T is the time required to pass over the route. I also assume a relationship among the variables. Although the steps and the length are assumed to be in linear dependence, the length and the time are assumed to be divided into two cases, one case constitutes a constant velocity walk, and the other case, a zero velocity walk, i.e., temporarily stopping during a walking task.

The equations below denote the relationship between L and W, and the relationship between T and L in the constant velocity walk.

$$L = \alpha \cdot W + \beta. \tag{3.1}$$

$$T = \gamma \cdot L. \tag{3.2}$$

α, β, and γ are the coefficient parameters for a subject as decided by the normal walking data in task 4. When the data sampled in short normal walking is denoted as $\{W_0, L_0, and T_0\}$ and that in long normal walking as $\{W_1, L_1, and T_1\}$, the coefficient parameters for the subject are denoted as below :

$$\alpha = \frac{L_1 - L_0}{W_1 - W_0}. \tag{3.3}$$

$$\beta = \frac{L_0 W_1 - L_1 W_0}{W_1 - W_0}. \tag{3.4}$$

$$\gamma = \frac{T_1 - T_0}{L_1 - L_0}. \tag{3.5}$$

When the sampled data in task 1 are denoted as $\{W_p, T_p\}$ and the data in task 3 as $\{W_s, T_s\}$, the length of the route the subject walked in each task is assumed using equations 3.1, 3.3 and 3.4 as below :

$$L_p = \alpha \cdot W_p + \beta. \tag{3.6}$$

$$L_s = \alpha \cdot W_s + \beta. \tag{3.7}$$

L_p is the full length of the route in task 1, and L_s is that in task 3.

The parameters for evaluating the contribution of the *I'm Here!* system consists of L_f, T_f and T_z. L_f is the length of additional roaming in task 3, as calculated using equations 3.6 and 3.7 as below :

$$L_f = L_s - L_p. \tag{3.8}$$

T_f is the time required by the additional roaming denoted using equations 3.2 and 3.5 as below :

$$T_f = \gamma \cdot L_f. \tag{3.9}$$

T_z is the time required by staying denoted using the result of equation 3.9 as below :

$$T_z = T_s - T_p - T_f. \tag{3.10}$$

Results

Table 3.1 Configuration of the number of trials in the object-recognition rate settings

settings	without-system	0%	33%	66%	100%
number of trials	3	3	3	3	2

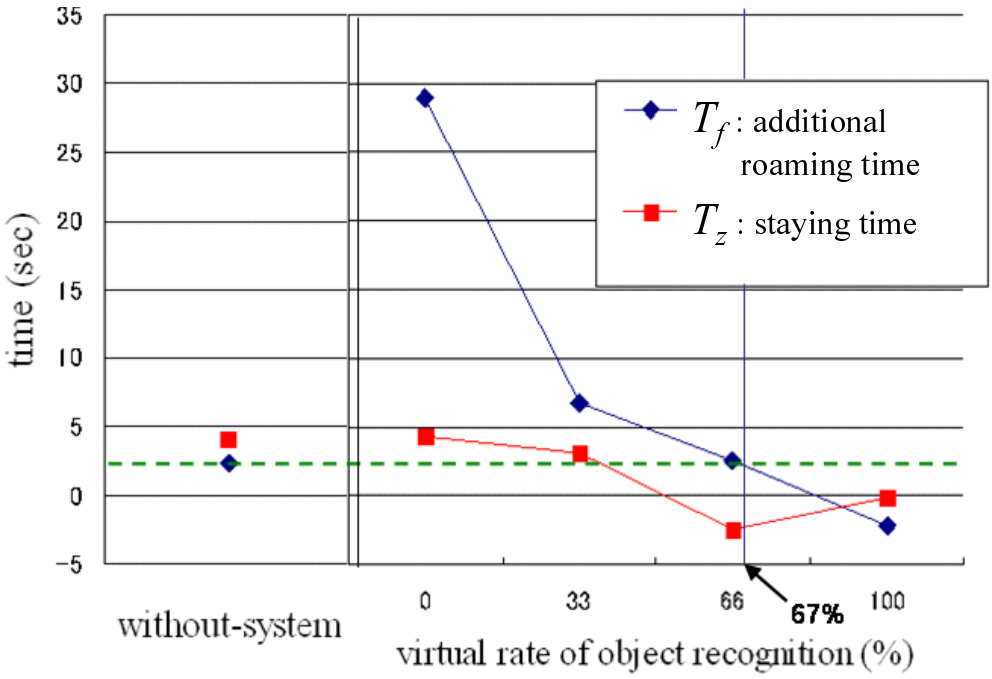

Figure 3.4 Contribution of the object-recognition rate to the object-finding task

Figure 3.4 denotes the experimental results evened off by the number of samples in each setting of the virtual object-recognition rate. Fourteen samples, which consist of the result of seven subjects each in two trials, are set as table 3.1. The horizontal dotted line denotes the score of T_f in the case that the subjects did not access the object retrieval function of the *I'm Here!* system. The line crosses to the result with the use of the system in about 67% of the recognition rate. The result implies that the system should embody a 67% accuracy at least.

Table 3.2 denotes the mean and standard deviation (S.D.) of T_f in each group of the cases in the trials of the object-finding task. The without-system case represents

Table 3.2. Means and Standard Deviations of the T_f

cases	without-system	negative	positive
N	9	18	15
Mean	2.3	19.5	-1.3
S.D.	5.1	35.2	1.4

the case without the support of the interface, and the positive/negative cases represent the cases of retrieved video that have true/false scenes. Via the result of ANOVA, I found the significant effect of providing the correct information of the indicated object for the time required by the subject's additional roaming in the object-finding task ($F_{(2,39)} = 3.41, p < .05$). The result of multiple comparisons based on the LSD method reveals that the mean of T_f in the negative case is significantly larger than in the positive case ($MS_e = 579.73, p < .05$). However, the result also reveals that the mean of T_f in the without-system case is not significantly different from that in the positive and negative cases.

Table 3.3. Means and Standard Deviations of the T_z

cases	without-system	negative	positive
N	9	18	15
Mean	-0.04	3.03	-0.29
S.D.	4.85	7.50	1.97

Table 3.3 denotes the mean and S.D. of T_z in each group of the trial case in the object-finding task. In contrast to T_f, no significant effect for the time required by temporarily stopping during a walking task from the result of ANOVA ($F_{(2,38)} = 1.56$) existed.

3.2.2 Discussions

On figure 3.4 I found that the *I'm Here!* system is effective with the user's object-finding task when the system has more than 67% of its object-recognition rate. Additionally, there appears to be a negative effect in using the system when the accuracy of

the system is below the rate. With the experimental results including interview logs, I analyzed noticeable cases where revealed worse results in the subjects' object-finding tasks with the system than the cases without the system. I found that the subjects in the worth cases had less reliability in their own memories. They blindly decided to rely on retrieved videos which actually represented negative cases. As the result of the decision, they were directed to where the target object was not placed, and roamed from there to find the target object.

On the interview logs, I found that subjects tended to evaluate the validity of a retrieved video by comparing episodes detected from the video with their own episodic memories. The episodes, for instance, if a user places a cup on a table, consist of the location of the cup on the table, other objects laid on there, appearance of the room from his/her viewpoint, a sequence of his/her activity, and so on. The episodic memories of the subjects more or less decreased with time.

I assume that a subject's reliance on his/her own episodic memories makes his/her evaluating validity of a retrieved video difficult. When a subject can not evaluate the validity, he/she has to make a choice between relying on the episodes detected from the video blindly and ignoring them. If he/she ignores them, the efficiency of his/her object-finding task will be the same as the task without the system. To make a user's evaluating validity of a retrieved video easier, extensional information, e.g. the time mark of the video, should be shown to him/her.

Additionally, I found that there were certain strategies applied to object-finding tasks by certain subjects who did not rely on the system. One of the strategies was to eliminate places where the target object would not be placed. Another strategy was to test the most possible candidate places where the target object would be placed. The enhancement of the *I'm Here!* system, e.g. displaying where the target object was often placed, will support these strategies and boost the efficiency of the user's object-finding task.

3.3 Reliability of *I'm Here!*

As I mentioned the last section, validity of a first-person video is one of the most important features for reliability of *I'm Here!*. A video that *I'm Here!* displays a user is constrained in its view size by hardware limitation of a head-mounted camera. Within design phase of *I'm Here!* system, the view size should be configured to an appropriate value. In this section, I explain a person's memory activity in recognizing contexts of the first-person video for the purpose of finding a target object. The explanation is based on my model of human memory activity compared to an electric circuit. By using the model to analyze results of experiments with subjects, I discuss an optimal view size condition of the first-person video to support human memory activity in object-finding task.

3.3.1 View size of video supporting object-finding task

I first design a model of a human memory activity related to an object-finding task. I 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, I 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.

Model of Human Memory to find object

Figure 3.5 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

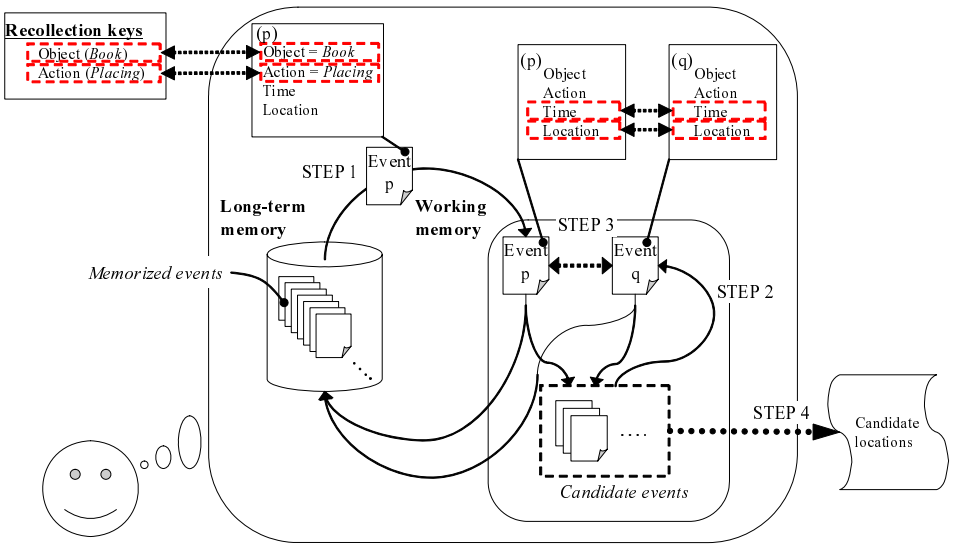

Figure 3.5. Model of human memory to find an object.

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.

I 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.

Effectiveness of first-person 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 3.5, will be reinforced by the results of (1) and (2) below :

(1) Newtson, D. A., (1976) notes that human can perceive a continuous action stream as a sequence of clearly segmented action units. When a user watching a first-person video perceives the action placing the target object, he/she should recognize any contexts in the video necessary for the perception. I assume that there is a key context called "action segmentation point." I also assume that the action segmentation point accompanies "preparative action," such as a movement of hand with the target object preparing to place. Figure illustrates the overview of them.

(2) When he/she watches a first-person 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. I 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.

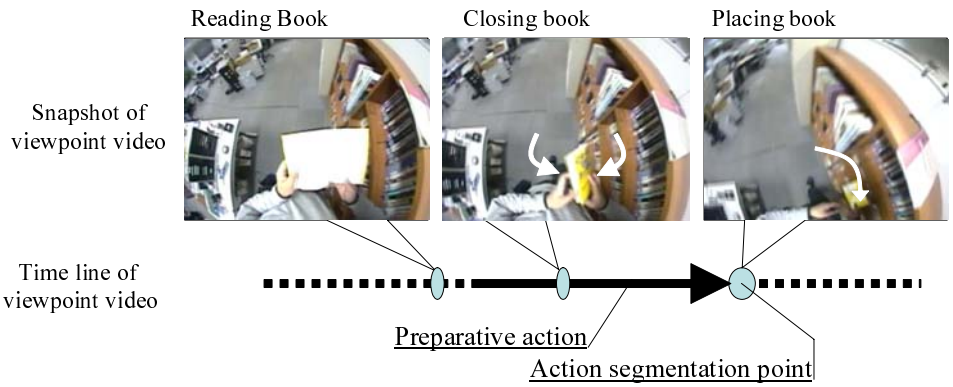

Figure 3.6. Overview of contexts for perceiving an action placing a book.

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).

Limitation of the first-person video

As I 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 size which causes problems such as when the user is unable to receive the benefits of the video and 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 I 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. I assume that the effectiveness of wider view angle is limited by a ceiling.

The conditions on the first-person video are inevitably constrained as below :

Limitation of view size : 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 , 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 size, 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 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 size should be investigated.

3.3.2 Experimental Evaluation

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

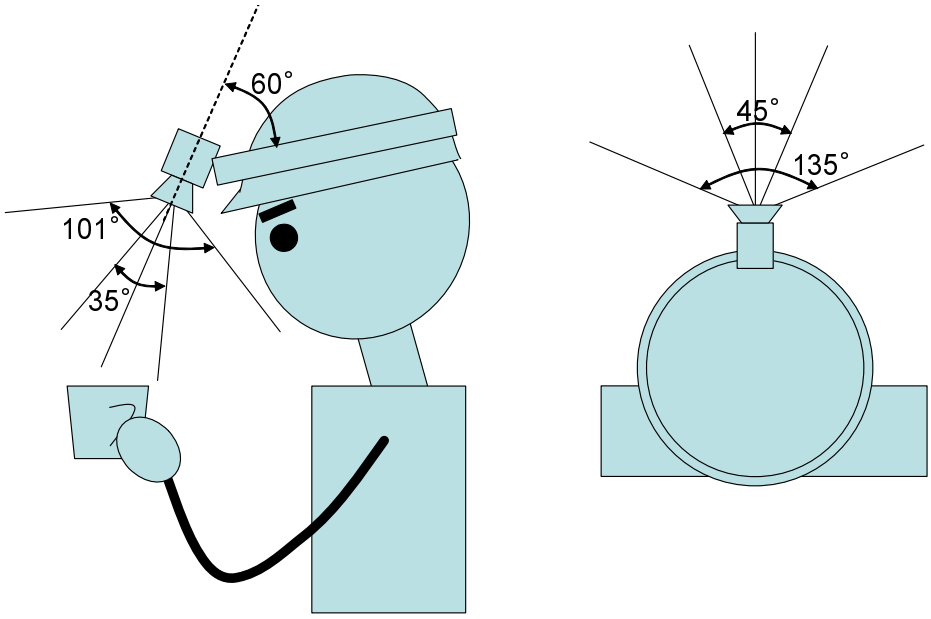

Figure 3.7. A camera device to record source videos.

Conditions

There were 15 subjects participating in these experiments. All of subjects were male students of graduate school studying in a laboratory. In an experimental trial, each subject watched a video displayed on a LCD to answer questionnaire about the context of the video.

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 worn by him (Figure 3.7). The camera was mounted on the front of his head with looking 60 degree of depression.

View size conditions of videos were virtually simulated with trimmed source videos (Figure 3.9). To establish 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 total 10 view size stages to simulate vies size conditions.

Each source video was recorded when the experimenter played a positive/negative action placing a target object on a place. The target object was only a book. There were 10 target places and all of them were parts of the laboratory in which subjects have spent usually (Figure 3.8). The positive action means that the experimenter actually

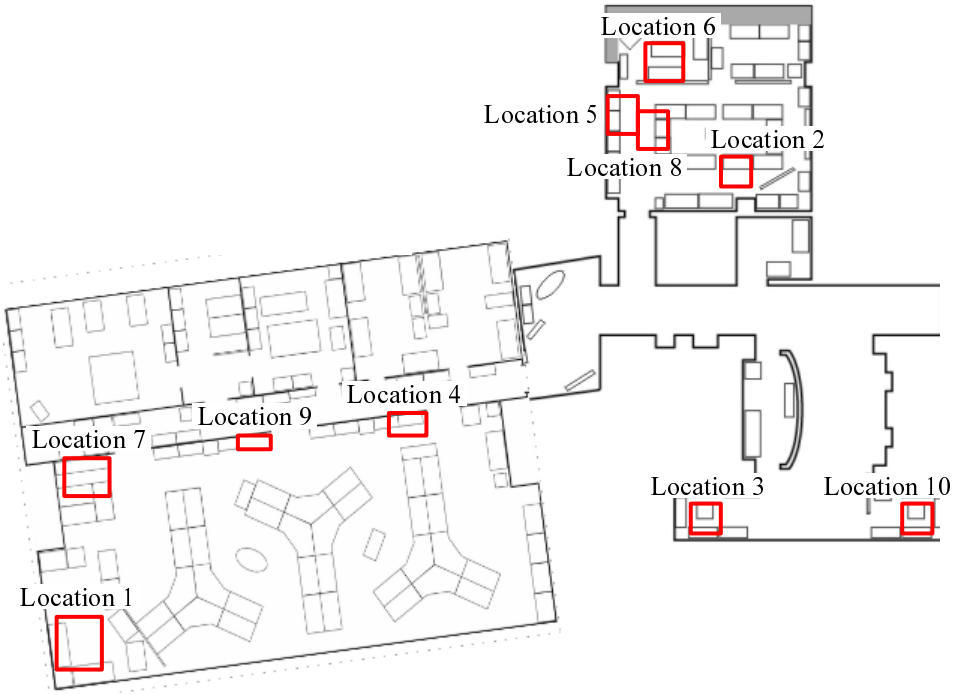

Figure 3.8. Laboratory environment and target locations.

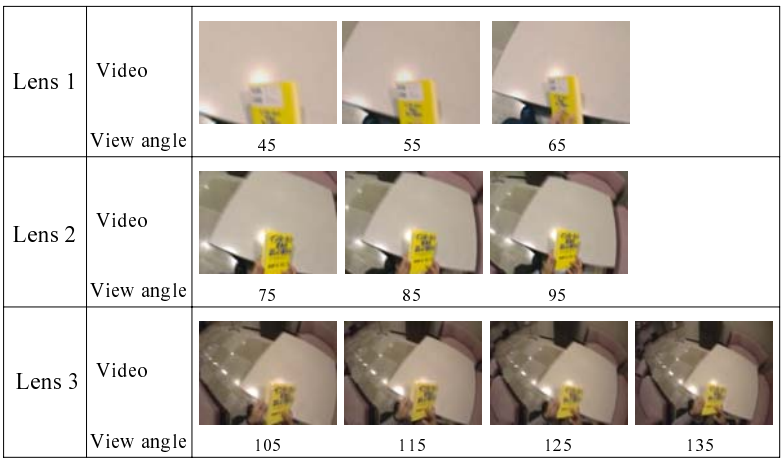

Figure 3.9. First-person video trimmed in view size stages.

Table 3.4. Action recognition patterns.

		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)

placed the target object on a target place. In contrast, the negative action means that he did not placed the target object on there but really kept holding it to bring to the other place. Positive actions were held in 5 places, half of all places, and negative actions were held in the other half. The experimenter played these behaviours naturally during recording videos.

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 10 view size groups consists of 10 videos of a common view size level with different 10 places. In case of focusing a lens used at recording, each lens group could be made up of videos extracted from a common source video recorded by the lens.

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

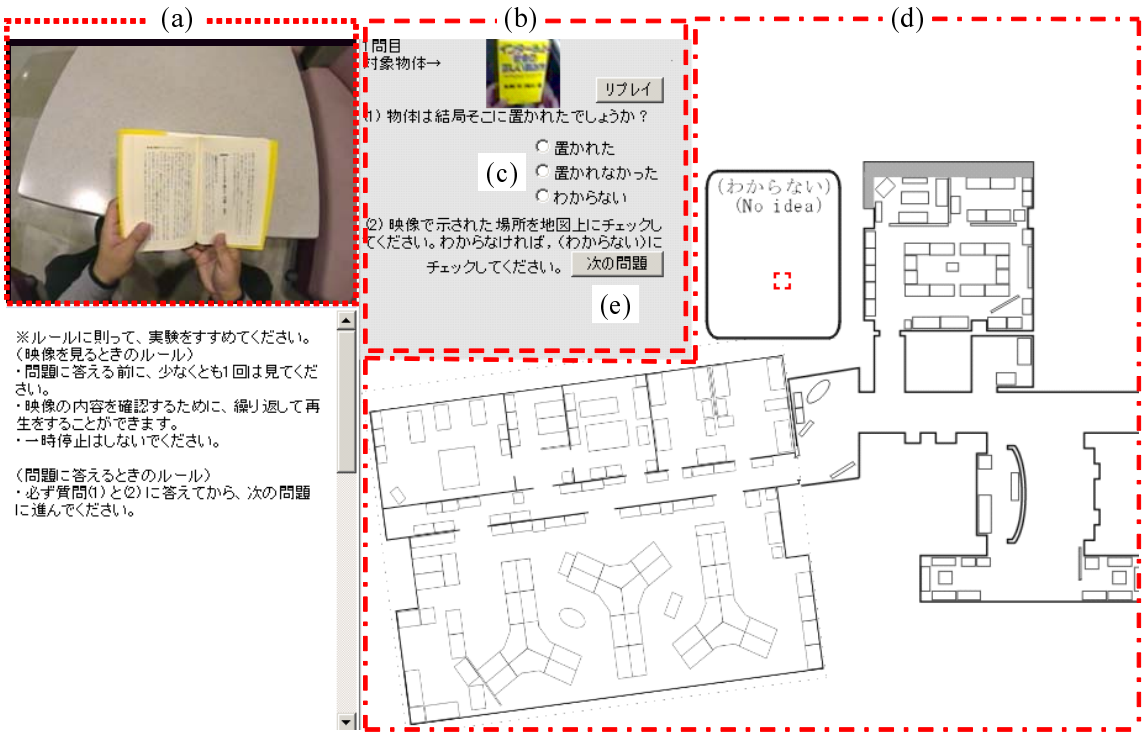

Figure 3.10. Voting Interface.

(a) video display area (b) questionnaire display area (c) action voting forms (d) place voting map (e) questionnaire feeding button

source videos :

1. The end-point was determined when any of following requirements was satisfied :

   (a) The case that occluded area of the target object became to be 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 case that visible areas of the target object became to be under 2% of the entire view area.

2. The start-point of experimental videos was determined with applying a 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 with voting interface. The voting interface illustrated in Figure 3.10 consists of video display area (Figure 3.10(a)), questionnaire display area (Figure 3.10(b)), action voting forms (Figure 3.10(c)), place voting map (Figure 3.10(d)) and a questionnaire feeding button (Figure 3.10(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 3.4.

The second question mentions cognitive estimations of each subject at viewing a place displayed in a first-person 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 size 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 size.

Results

From questionnaires written by subjects, results of action recognition and location recognition were counted on view size 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 size conditions, generated from 4 sources of positive case and 4 of negative case, were totally counted.

Figure 3.11 illustrates subject number ratio of action recognition separated into TRUE-POSITIVE (Figure 3.11 (a)), FALSE-NEGATIVE (Figure 3.11 (b)), FALSE-POSITIVE (Figure 3.11 (c)) and TRUE-NEGATIVE (Figure 3.11 (d)). 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

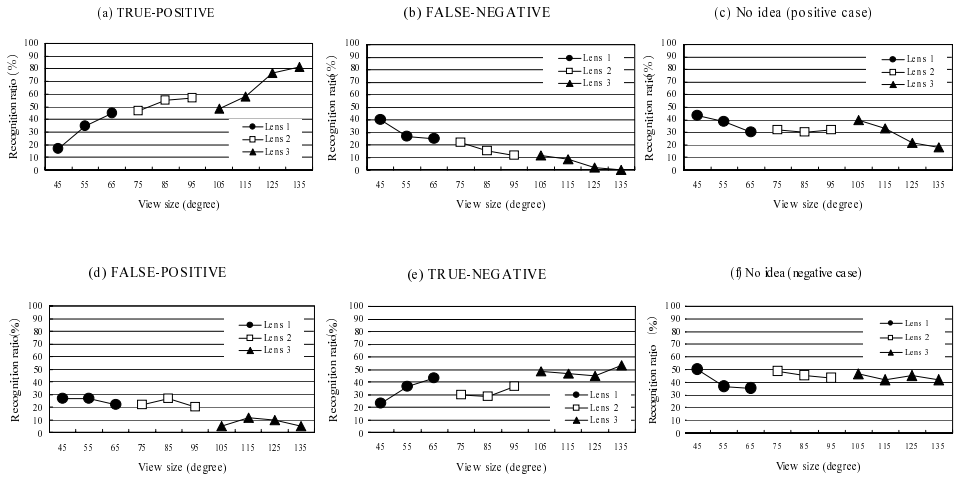

Figure 3.11. Result of action recognition

whether the action was actually occurred / not occurred. I found that the ratio of correct answer increases, and that of incorrect answer decreases with increasing view size condition.

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 3.12, the number of candidate locations was well narrowed down by setting view size condition to 115. Figure 3.13 (a) illustrates that the true results of location recognition were saturated from 125 of view size condition. On the same setting, false graph (Figure 3.13 (b)) and no idea graph (Figure 3.13 (c)) were asymptotically stable at each rock-bottom value.

3.3.3 Discussion

With experimental results of location recognition, I define required view size conditions as 115 to support a user 's object-finding task. Figure 3.13 (b) illustrates, however, that some subjects answered wrong locations on the view size condition.

It is an important approach to extend the time length of a video for preventing

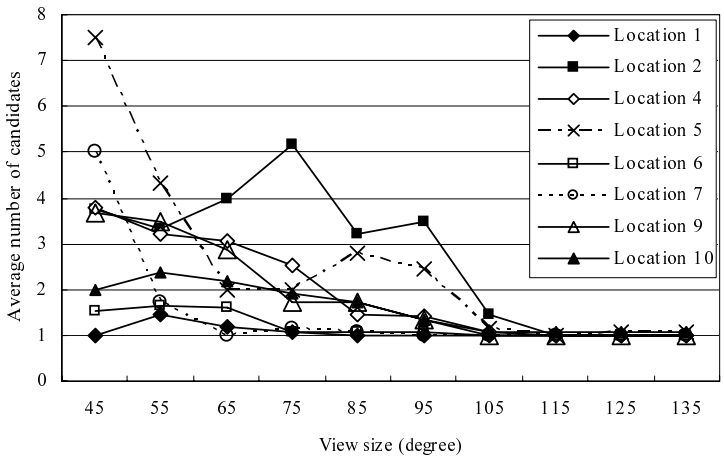

Figure 3.12. Narrowing down of candidate locations.

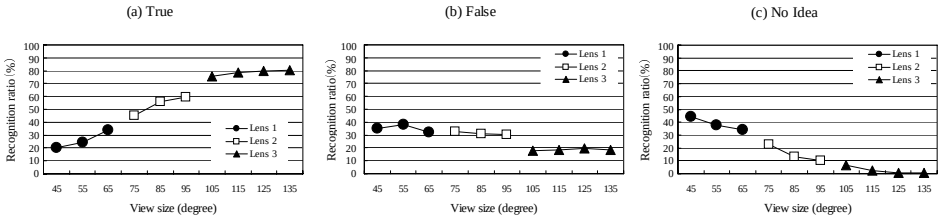

Figure 3.13. Result of location recognition

such misunderstandings of location recognition. If the time length of a video will be extended to back and forth, a user watching the video will be able to recognize other locations nearby where the target object was placed. He/she then can choose a plan to search for the object with walking through those locations.

On the other hand, the sufficient view size 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 size 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!*.

I found that subjects estimated the action placing the target object with its preparative action, even if they could not recognize the action segmentation point. The gradual expansion of the TRUE-POSITIVE graph (Figure 3.11 (a)) supports the presence of the estimation.

The action estimation, at the same time, causes the action recognition failure of FALSE-POSITIVE pattern. I 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. To constrain the pseudo action segmentation, *I'm Here!* should display any additional first-person 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.

Chapter 4

I'm Here! Implementation

This chapter explains about an implementation and evaluation of camera devices and an object recognition method which are designed to be used in *I'm Here!*

I'm Here! is a wearable system for indexing / retrieving Experience Database. The Experience Database is a database of a first-person video indexed by names of objects registered in Object Database. The accuracy in indexing the Experience Database is important for effectiveness of *I'm Here!*.

Above all functions in *I'm Here!* implementation, the function to construct Experience Database has to be preferentially discussed. The function consists of an object image extraction module and an vision-based object identification module. I developed a wearable camera ObjectCam to achieve real-time object extraction. In this chapter I discuss a wearable camera ObjectCam and ObjectCam2 to be utilized in everyday environment. I also explain the vision-based object identification module for indexing the Experience Database.

4.1 Overview of *I'm Here!* Implementation

I'm Here! is composed of a PC, a Head-Mounted Display (HMD), a wearable camera ObjectCam, and an JogDial interface. Figure 4.1 is hardware construction of *I'm Here!*. The *ObjectCam* captures a first-person video. The HMD shows the user system information and the retrieved video. The JogDial interface is used to select an item displayed in the HMD.

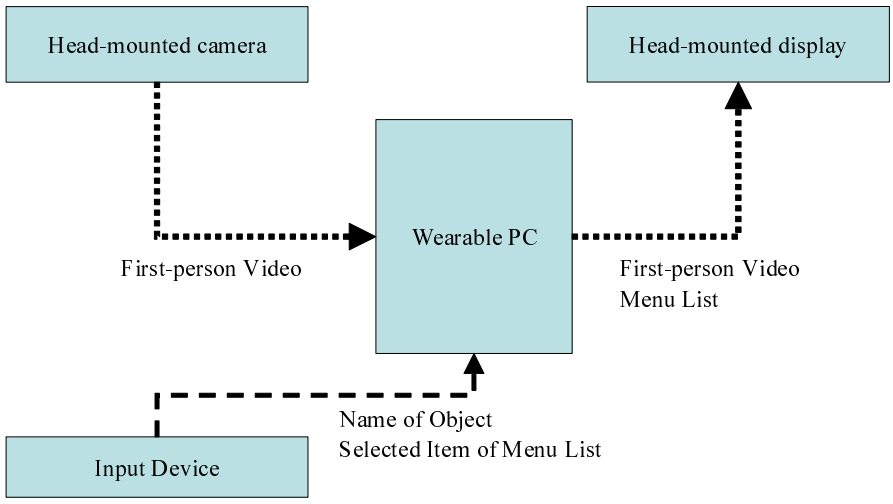

Figure 4.1. Device construction of *I'm Here!*

4.2 Object Extraction and Recognition

4.2.1 ObjectCam

The image of a user's viewpoint includes the object region, the region of the hand holding the object, and the background region. To construct an appearance-based image-feature of the object, the region of the object must be extracted from the image by eliminating both the hand region and the background region.

Many studies of extracting a target region by eliminating a background region exist. The background subtraction method has been used in many real-time vision systems with fairly good results (Wren et al. (1997)). However, this method requires a static background image that can rarely be obtained by a wearable camera. On the other hand, top-down knowledge from object recognition that can be used for segmenting the target region from exterior regions has also been discussed (Leibe et al. (2003)). This method is suitable for images with cluttered backgrounds and partial occlusions. However, applying this method to real-time applications is difficult because the temporal loads from preliminary learning of knowledge and iterative processing are

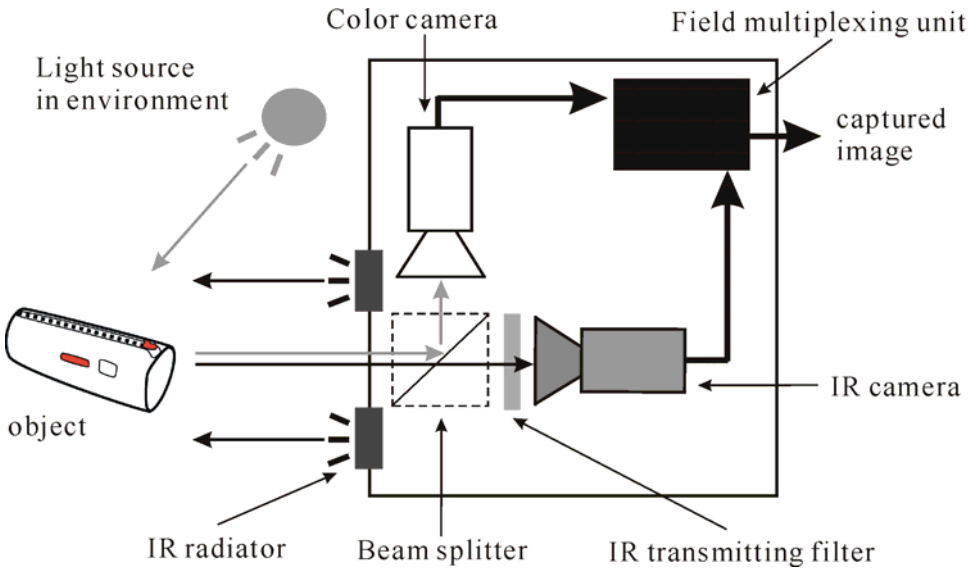

Figure 4.2. Architecture of ObjectCam

required in this method.

I have developed a new camera device named "ObjectCam" to extract only the object image from the user's viewpoint image. This camera device enables low-cost object image extraction regardless of background complications.

Figure 4.2 illustrates the architecture of the "*ObjectCam*." The camera device consists of a color camera and an infrared camera clamped at the same posture across a beam splitter. A frame of the image captured by the *ObjectCam* consists of a color and an infra-red (IR) image for each field. An IR image displays the reflected IR luminance caused by the IR radiator on the front of the device.

Figure 4.3 illustrates a blockgram of the object extraction process. First, a nearby region mask is made from an IR image with a luminance threshold in a binarizing process (figure 4.3(a)). Second, by applying the mask to remove the background region from the color image (figure 4.3(b)), the system creates a nearby region image. Third, the object region mask is created by removing the region of the hand holding the object using the user's skin color (Bergener, T. et al. (1997)) (figure 4.3(c)). Finally, by applying the object region mask to each color and each IR image of the user's viewpoint, the system creates each image of the object (figure 4.3(d)).

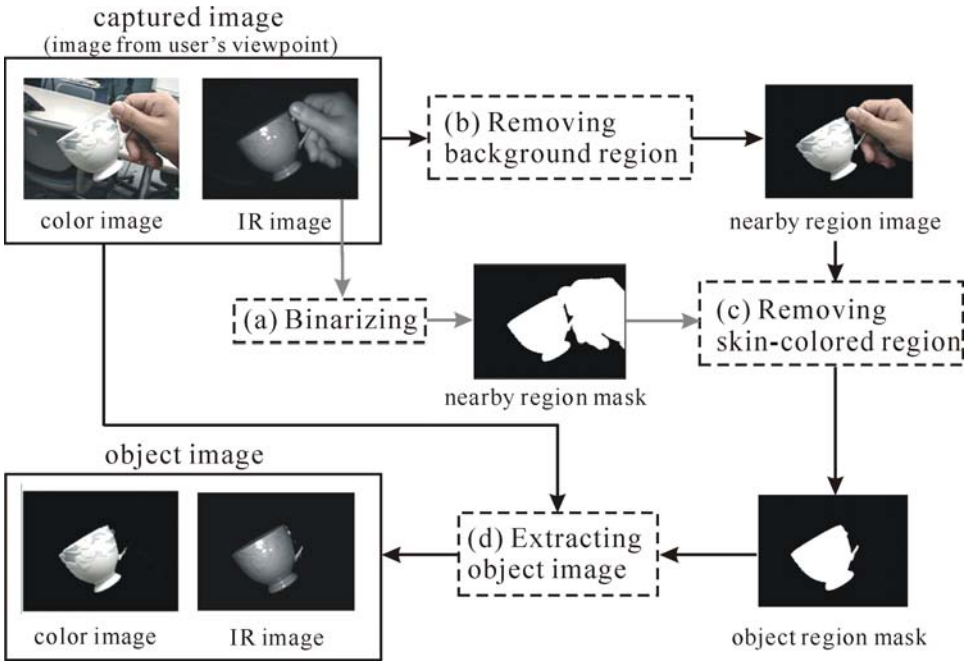

Figure 4.3. Object extraction by ObjectCam

4.2.2 Construction of Object Database

Both object registration and recognition processes are necessary for retrieving the last recorded video containing the target object from the video memory database. Many object recognition methods have been proposed. The appearance matching of three-dimensional objects using parametric eigenspace features, for instance, has been estimated with good results (Schiele et al. 1996). However, a large load is necessary to compress large amounts of input images into the eigenspace features, and a strict facility to register the eigenspace features is required. On the other hand, probabilistic object recognition using the feature of multidimensional histograms has been applied to estimate the probabilistic presence of the registered objects in the scene (Nayar et al. (1996)). Because the experimental result mentions only the scale change and image-plane rotation of the input image, the variety of appearances of the three-dimensional object still remains a considerable problem.

Several problems regarding the recognition of the object held by a user in everyday circumstances still need to be considered; for example, 1) real-time registration and recognition of the object, 2) the ability to recognize a three-dimensional object in several appearances, and 3) the problem of achieving good performance of object recognition

in the case of increasing the amount of registered objects. The study proposing an appearance matching of three-dimensional objects using parametric eigenspace features has the disadvantage of 1), and the other study of probabilistic object recognition using the feature of multidimensional histograms does not clearly address 2). In the development of "*I'm Here!*", a multi-dimensional histogram feature extracted from object images captured by *ObjectCam*, which includes color and IR luminance data is proposed to solve problems 1) and 2), and the proposed feature is estimated from the perspective of 3) in the section, "Experiments and Result."

The object dictionary contains the appearance-based image-features of the registered objects with their IDs. An object-feature of a three-dimensional object is constructed from several images in representative appearances. To construct the object-feature, the system captures several images of the object, makes image-features representing each image, and integrates these image-features into an object-feature by grouping and integrating similar image-features.

An object image-feature is denoted by a $\{H\text{–}Z\text{–}C\}$ three-dimensional histogram. This histogram consists of $\{H\text{–}Z\text{–}C\}$ elements extracted from each pixel of the object image. H and Z represent Hue and IR luminance value. C represents the pixel group ID divided by distance from the centroid of the silhouette of the object image. In addition to the hue values as a color feature, the $\{H\text{–}Z\text{–}C\}$ histogram includes IR luminance as a depth-like feature and as a silhouette-like feature. Due to each feature's robustness for the rotation of the view axis and the low-cost processing of this histogram feature, the $\{H\text{–}Z\text{–}C\}$ histogram is expected to be a good image-feature in object recognition.

The hue value is based on HSV color representation converted from the RGB values of a pixel. The system extracts hue value H using the expressions below :

$$V = \max(R, G, B) \tag{4.1}$$

$$W = \min(R, G, B) \tag{4.2}$$

$$S = \alpha\left(\frac{V - Z}{V}\right) \tag{4.3}$$

$$H = \begin{cases} \beta\left(\dfrac{G - B}{V - Z}\right) & , \quad R \equiv V \\ \beta\left(2 + \dfrac{B - R}{V - Z}\right) & , \quad G \equiv V \\ \beta\left(4 + \dfrac{R - G}{V - Z}\right) & , \quad B \equiv V \end{cases} \tag{4.4}$$

S represents the saturation value, and V and W represent the maximal and minimal value among the $\{R,\ G,\ B\}$ values of the pixel. These values are limited as $0 \leq$

$R, G, B, S, V, W < \alpha$ and $0 \leq H < \beta$.

The hue value with low saturation is sensitive to sensor noise, however. According to the sensor noise, the and color features oscillate on their own order. Furthermore, the hue value becomes uncertain as the saturation value decreases. This creates a problem.

To avoid the influence of the sensor noise I apply the probabilistic sensor noise model to the hue value. The distribution of sensor noise is assumed to the Gaussian distribution, as denoted below when the average of the distribution is u and the dispersion is σ^2 :

$$K(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-u)^2}{2\sigma^2}}. \tag{4.5}$$

Using equation 4.4, the dispersion of H is represented as below :

$$\sigma_H = \begin{cases} \sigma_{RGB}(G,B) & , \quad R \equiv V \\ \sigma_{RGB}(B,R) & , \quad G \equiv V \\ \sigma_{RGB}(R,G) & , \quad B \equiv V \end{cases} \tag{4.6}$$

The element σ_{RGB} in equation 4.6 is represented as a follow :

$$\sigma_{RGB}(x,y) = \alpha\beta\sqrt{\frac{(x-y)^2(V^2\sigma_S^2 + S^2\sigma_V^2) + S^2V^2(\sigma_x^2 + \sigma_y^2)}{V^4S^4}}. \tag{4.7}$$

The dispersion of the $\{S,\ V,\ W\}$ value represented as $\{\sigma_S, \sigma_V, \sigma_W\}$ can be directly derived from the statistical observation of the $\{R,\ G,\ B\}$ distribution.

Using equation 4.5, 4.6, and 4.7, the uncertainty of the hue value with low saturation can be represented; in the case of zero saturation, the probabilistic hue distribution becomes flat.

An image-feature of an object is the integrated distribution of all the pixels of the image represented as the $\{H\text{–}Z\text{–}C\}$ histogram feature. Figure 4.4 represents the construction of the $\{H\text{–}Z\text{–}C\}$ histogram feature from an object image when th pixel of an object image has the hue value of H_i, the IR luminance value of Z_i, the group ID of C_i, and the distance from the centroid of the silhouette of the object image of L_i. The value of C_i is denoted as below :

$$C_i = \left[\frac{L_i}{L} \cdot \sqrt{\frac{n_0}{n}}\right]. \tag{4.8}$$

L represents a standard distance. n equals the entire amount of all pixels of the object image and n_0 represents the normalized amount of pixels. As depicted in Figure

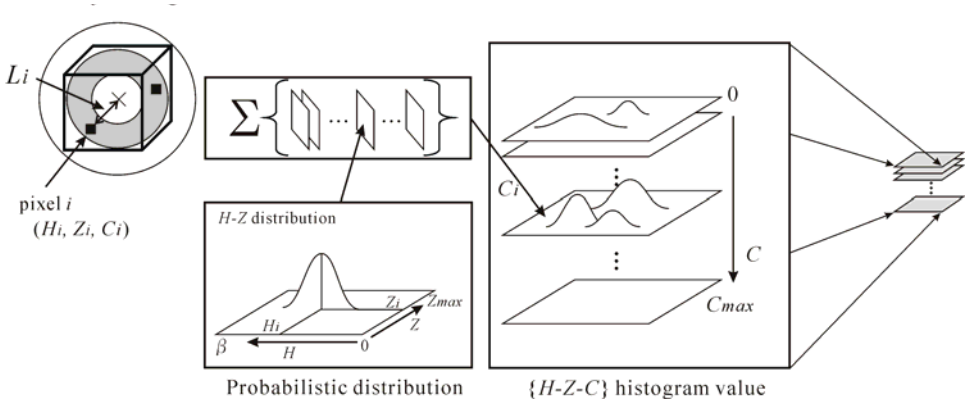

Figure 4.4. $\{H\text{-}Z\text{-}C\}$ histogram

4.4, the distribution has a hue axis, an IR luminance axis, and an axis of the pixel group number. The values of Z and C are limited as $0 \leq Z \leq Z_{\max}$ and $0 \leq C \leq C_{\max}$. The distribution value in $\{H = j, Z = k, C = l\}$ is represented as follows :

$$D(j,k,l) = \sum_{i=1}^{n} f_H(i,j) \cdot Z_{ik} \cdot C_{il}, \tag{4.9}$$

$$Z_{ik} = \begin{cases} 1, & k = Z_i \\ 0, & k \neq Z_i \end{cases}, \tag{4.10}$$

$$C_{il} = \begin{cases} 1, & l = C_i \\ 0, & l \neq C_i \end{cases}, \tag{4.11}$$

$f_H(i,j)$ represents a probabilistic contribution of ith pixel for the jth bin of the hue value represented as a follow :

$$f_H(i,j) = \frac{1}{n} \int_{h(j-1)}^{hj} K_i^H(x)dx, \{j | j = 1, 2, ..., \frac{\beta}{h}\}. \tag{4.12}$$

h is the breadth of the bin of the hue (Kawamura, T. et al. (2003)).

From several appearances of an object, the system constructs an object-feature as a set of selected representative image-features (Figure 4.5). The selection of representative image-features is based on a grouping of similar image-features of an object.

The system captures several images of the object and calculates the $\{H\text{–}Z\text{–}C\}$ features of each image. I can reduce the amount of the elements of an object-feature by grouping similar $\{H\text{–}Z\text{–}C\}$ features into a representative image-feature to avoid a redundant comparison in identifying the object.

The similarity of image-features D_i and D_j is based on the Sum of Absolute Difference (SAD) of three-dimensional histograms, as denoted by $\mathbf{SAD}^*(D_i, D_j)$ in the

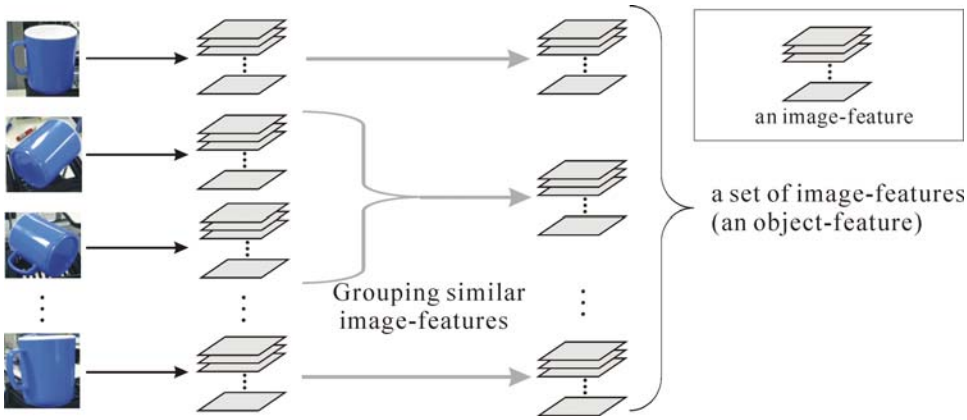

Figure 4.5. Construction of object-feature

equations below :

$$\mathbf{SAD}^*(D_i, D_j) = \begin{cases} \sum_{h,z,c} |D_i(h,z,c) - D_j(h,(z+z_m(i)-z_m(j)),c)|, & 0 \le (z+z_m(i)-z_m(j)) \le z_{\max}, \\ \sum_{h,z,c} |D_i(h,z,c)|, & other. \end{cases} \tag{4.13}$$

$z_{\max}$ represents the maximal value of the IR luminance values. When $D_x(h,z,c)$ is reintegrated into $D_x^Z(z)$ on the basis of IR luminance, $z_m(x)$ represents the centroid of $D_x^Z(z)$. $z_m(x)$ and $D_x^Z(z)$ are denoted as follows :

$$z_m(x) = \frac{1}{n}\sum_z D_x^Z(z), \tag{4.14}$$

$$D_x^Z(z) = \sum_{h,c} D_x(h,z,c) \tag{4.15}$$

The $\mathbf{SAD}^*$ is derived by aligning distributions with their centroid on the basis of IR luminance values, because the reflected IR luminance value is experimentally observed in linear relation to the distance from the object's surface.

The following equations define the clustered structure of the images of the object :

$$\begin{aligned} \mathbf{A}^{U_j} &= \{A | 1 \le i \le N_j\}, \\ \mathbf{A}^{U_1} &= \{\mathbf{Q}_i | 1 \le i \le N_1\}, \\ \mathbf{A}^{U_{(j+1)}} &= A^{U_j} - \mathbf{Q}_j. \end{aligned} \tag{4.16}$$

$\mathbf{A}^{U_j}$ represents the jth set of images of the object. $\mathbf{A}^{U_j}$ includes all images of the object. A_{ij} represents the ith image included in $\mathbf{A}^{U_j}$. N_j represents the amount of images included in $\mathbf{A}^{U_j}$, and is bounded in $1 \le N_j \le N_1$. $\mathbf{Q}_i$ represents the ith cluster

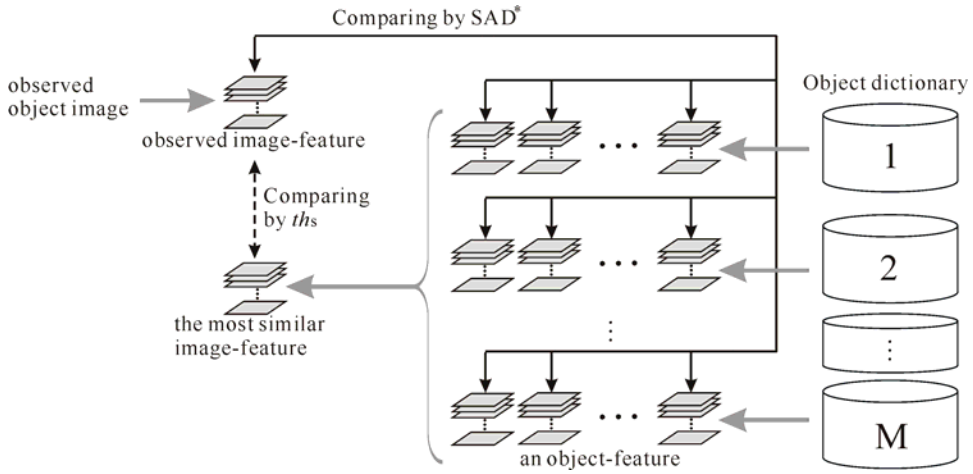

Figure 4.6. Matching of object-feature

of similar images included in $\mathbf{A}^{U_j}$. When R_j is the selected representative image of $\mathbf{A}^{U_j}$, R_j is denoted as follows :

$$R_j = A_{1j}. \tag{4.17}$$

When the image-feature of A_{kj} is D_{kj} and the image-feature of a representative image R_j is D_{1j}, the following equations define the contents of $\mathbf{Q}_j$ with equation 4.13 :

$$\begin{cases} A_{kj} \in \mathbf{Q}_j, & \mathbf{SAD}^*(D_{1j}, D_{kj}) < th_q, \\ A_{kj} \notin \mathbf{Q}_j, & other. \end{cases} . \tag{4.18}$$

th_q represents the threshold for $\mathbf{Q}_j$ to content images similar to R_j. th_q should be defined by an empirical method so that images of an object with slightly shifted appearances can be classified in a same image group. The set of selected representative images $\mathbf{R}$, is defined as a follow :

$$\mathbf{R} = \{R_i | 1 \leq i \leq N_R\}. \tag{4.19}$$

N_R represents the amount of selected representative images of the object bounded in $1 \leq N_R \leq N_1$.

In the object observation phase the system tries to recognize an observed object along with any of the registered objects. In the recognition process, the observed object-feature is compared with all registered object-features in terms of similarity (Figure 4.6).

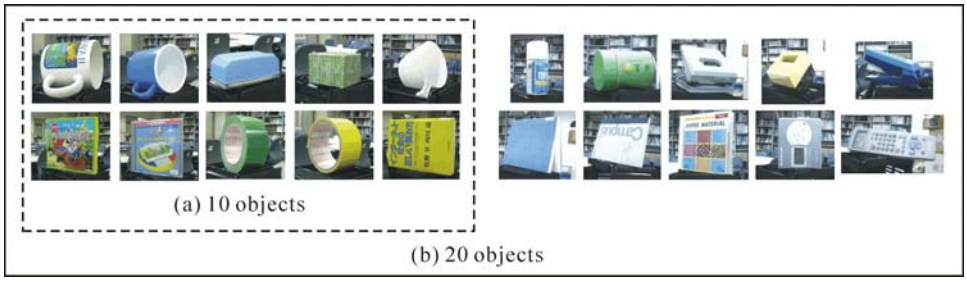

Figure 4.7. The sets of target objects

I defined the similarity between observed object o and a registered object r using equation 4.13 as below :

$$e_r = \min\{\mathbf{SAD}^*(D_o, D_p)|p = 1, 2, ..., N_r\}. \tag{4.20}$$

D_o represents the image-feature of the observed object o, and D_p represents the pth image-feature of the registered object r. N_r represents the amount of selected representative images of object r.

The result of object recognition is represented as below :

$$Obj = \begin{cases} r, & \min\{e_r|r = 1, 2, ..., M\} < th_s, \\ 0, & other. \end{cases} . \tag{4.21}$$

M represents the amount of registered objects, and th_s the threshold of minimal similarity. The result 0 shows that there are no registered objects matched with the observed object. th_s should be defined by an empirical method so that the result of object recognition with any unregistered object can be 0.

Through the object observation phase, the system creates the video memory, which is the video data with the index of observed objects. The index is associated with every frame of the video memory. If a frame of captured video contains a registered object, the index explicitly annotates to the frame with the ID of the object.

4.2.3 Experimental Evaluations of Object Identification

Since the object recognition performance affects the accuracy of retrieved video, I have estimated the object recognition performance of the proposed method by a few experiments. These experiments were performed in an indoor environment with fluorescent light. The system is estimated in offline usage, by setting ObjectCam on its base and by setting a target object on a turntable.

Table 4.1. Experimental results

Object feature value	H	H-C	H-Z	H-Z-C
(a) 10 objects	99.2	99.2	91.7	96.7
(b) 20 objects	81.7	88.8	87.5	94.1

The sets of target objects are depicted in Figure 4.7. Set (a) consists of ten objects, and set (b) consists of twenty objects including objects in (a) and an additional ten objects. Each object is rigid, the appropriate size to hold, and used in everyday life.

The target objects have been utilized in the everyday environment for students. They are selected with rules below :

In offline object registration, the system captures twenty images for each object as input images for object registration. Each of the twenty images is in a different configuration of distance and perspective. The system creates an object dictionary with selected representative image-features and an ID of each object.

Instead of the object observation phase, the system calculates an object recognition rate with every registered image to be observed, i.e., the object dictionary consists of all the registered object-features in these experiments. The system rejects matching between different objects and allows matching between any configuration patterns in the same object.

To demonstrate the advantage of the proposed method, I have compared the $\{H$–Z–$C\}$ method with other methods using $\{H\}$, $\{H$–$Z\}$, and $\{H$–$C\}$ features in these experiments. The parameters are set as follows: $\alpha = 256, \beta = 360, Z_{\max} = 255, C_{\max} = 10, L = 10, n_0 = 1000$ and $h = 4$. Table 4.1 displays the result of the experiments.

4.2.4 Discussion

I found that in the proposed method, the recognition rate decreases less than in other methods even when the amount of objects in the target set increases. In everyday use of "*I'm Here!*", the amount of registered objects increases. The experimental result demonstrates that the proposed method is useful for online object recognition in "*I'm Here!*"

4.3 Environmental Adaptation

Above all devices constructing I'm Here, ObjectCam should be minitualized in preference. Size and weight of head-worn devices, a head-mounted camera and head-mounted display, mostly affect barycentric position of entire wearable system. The minitualization of ObjectCam efficiently improves the comfort of *I'm Here!* system for the user.

object-finding tasks can happen at various locations in everyday environment. There are many locations with sunlight illumination including IR light. The ObjectCam can be used only in the environment without IR illumination, such as indoor environment with fluorescent illumination after sunset. In the location with environmental IR illumination, the ObjectCam extracts not only nearby object image but also background object.

I challenged to solve both minitualization problem and the environmentally-dependent problem with developing ObjectCam2, an improved version of ObjectCam.

4.3.1 ObjectCam2

This section shows the features of *ObjectCam2* customized for identifying a trigger object under everyday conditions. The *ObjectCam2* has the following features :

- small and lightweight
- dividing a nearby object image from its background by 30Hz
- employing color and active IR images
- controlling blinking IR LEDs
- controlling the exposure time to respond to the problem of dropping out the highlighted surface of the nearby object

I certify the specifications of the *ObjectCam2* and describe a vision-based method for extracting the object image robustly under the everyday illumination conditions affected by sunlight.

Table 4.2. Specification of *ObjectCam2* and *ObjectCam*

	ObjectCam2	*ObjectCam*
Weight of camera head (g)	390	670
Size of camera head (mm) (width×height×depth)	95 × 70 × 90	80 × 170 × 120
Photo acceptance unit	1 CMOS (color and IR)	2 CCD (1 color and 1 IR)
Exposure time control	Yes	No
Output image	color and IR (full view) or (nearby object image)	color and IR (full view)
Shutter frequency (Hz)	30	30
Number of IR LEDs	32	98
Blinking control	Yes	No

Figure 4.8. A wearable camera *ObjectCam2*

Table 4.2 indicates the specifications of the *ObjectCam2*. The table also indicates the specifications of *ObjectCam*. Each camera is a kind of active IR camera with an IR LED array in its front. Each camera is used as a head-mounted camera mounted on the front of a helmet. Figure 4.8 indicates a scene using the *ObjectCam2*. The *ObjectCam2* is more miniaturized and made to be more lightweight than the *ObjectCam* so as to be employed as a wearable camera device. As opposed to *ObjectCam*'s continuous lighting of the IR LED array, *ObjectCam2* controls blinking in the IR LED array to capture a color image with reflected active IR light and to capture a color image without the lights. The CMOS unit of the *ObjectCam2* has no IR eliminating filter so that *ObjectCam2* can

capture an image that has luminance within the color and the IR light range. Using the images, *ObjectCam2* can divide a nearby object image from its background by 30Hz.

ObjectCam2 is suited to be utilized as a wearable camera. ObjectCam2 has been minitualized to be smaller and lighter than ObjectCam. The minitualization has decreased the number of LEDs arrayed on the front of ObjectCam2 into 32 from that of ObjectCam, 96. I have offseted the paucity of LED in ObjectCam2 with employing brighter LED than that employed in ObjectCam.

The *ObjectCam2* realizes the two functions shown below to identify a trigger object. 1) The function for employing multiple features of the object including color and active IR images (Mihara et al. (2002)). 2) The function of blinking IR LEDs for extracting a region of the nearby target image, including the image of the trigger object itself (Lee et al. (2002)). By integrating these two functions, *ObjectCam2* has an advantage among the other active IR cameras in extracting and recognizing trigger objects in the everyday environments of its user.

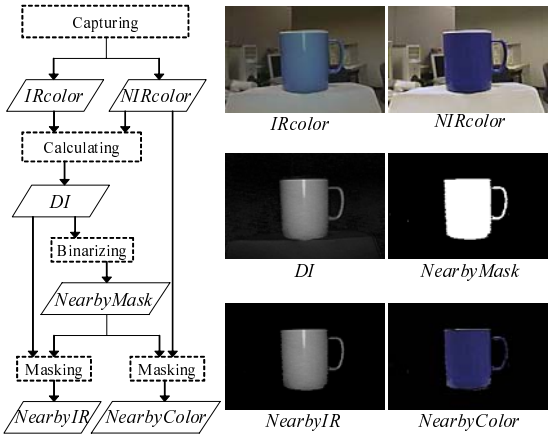

Figure 4.9. Process diagram for extracting a nearby target object

Figure 4.9 indicates a diagram of the integrated process for extracting the nearby target image. The $IRcolor$ is a color image with a reflected active IR light and the $NIRcolor$ is a color image without the light. The *ObjectCam2* captures the $IRcolor$ and the $NIRcolor$ by turns. In the equation below, *ObjectCam2* subtracts the IR luminance caused by environment IR illumination in both $IRcolor$ and $NIRcolor$:

$$DI = Y_{IRcolor} - Y_{NIRcolor}. \tag{4.22}$$

DI is an active IR image indicating the IR luminance only illuminated by active IR light. $Y_{IRcolor}$ and $Y_{NIRcolor}$ denote images only indicating luminance(Y) in YUV color feature of $IRcolor$ and $NIRcolor$. A mask image $NearbyMask$ is simply created by binarizing DI. The *ObjectCam2* finally creates an active IR image $NearbyIR$ and a color image $NearbyColor$ from DI and $NIRcolor$ by simply masking with $NearbyMask$. Extracted region of nearby target images are illustrated in $NearbyIR$ and $NearbyColor$.

I have solved the problem of the highlighted area dropping out from nearby object image extracted by the *ObjectCam2*. The problem is caused by luminance saturation in $IRcolor$. When the exposure time of $IRcolor$ and $NIRcolor$ are the same without being controlled, the luminance of $IRcolor$ is larger than that of $NIRcolor$ due to the active IR illumination. The luminance of the highlighted part of an object, such as the white-colored surface, tends to be saturated in $IRcolor$. The luminance of the part in DI is calculated to a smaller value than the proper value.

To avoid the saturation of $IRcolor$, the *ObjectCam2* controls exposure time at the time of capturing $IRcolor$. When $\delta_{IRcolor}$ and $\delta_{NIRcolor}$ indicate the exposure time of $IRcolor$ and $NIRcolor$, equation (4.22) should be fixed as below :

$$DI = \frac{\delta_{NIRcolor}}{\delta_{IRcolor}} Y_{IRcolor} - Y_{NIRcolor}. \tag{4.23}$$

The illumination condition of $Y_{IRcolor}$ and $Y_{NIRcolor}$ has been coordinated with a common environment illumination condition to calculate proper DI.

Figure 4.10 illustrates the result of controlling the exposure time. Without controlling the exposure time, only the luminance of the white-colored right-hand surface of the cube is calculated as smaller in DI, and the surface has been lost in $NearbyColor$. The *ObjectCam2* controls $\delta_{IRcolor}$ to a half of $\delta_{NIRcolor}$, and avoids dropping the part from the nearby object image with proper DI created from unsaturated $IRcolor$ and $NIRcolor$.

4.3.2 Experimental Evaluations of Object Image Extraction

I have evaluated the extraction accuracy of a nearby object image by *ObjectCam2* in experiments showing that *ObjectCam2* is suited for use in an everyday environment.

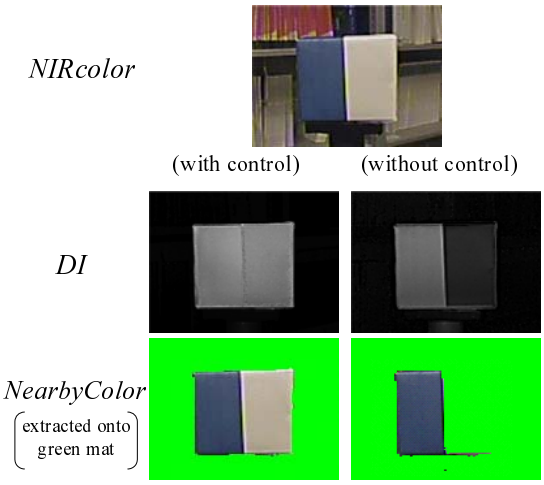

Figure 4.10. Result of controlling exposure time

The extraction accuracy of a nearby object image is indicated in the $NearbyMask$. By comparing the quality of the $NearbyMask$ created by *ObjectCam2* with that of the $NearbyMask$ created by *ObjectCam*, I have demonstrated the functional effects of the *ObjectCam*.

Experiments

Experiments were carried out at following four locations :

(a) Lobby with window facing to north

(b) The lobby facing to east (indoor background)

(c) The lobby facing to west (outdoor background)

(d) Laboratory at night without sunlight illumination

In each case of (a) (b) and (c), three experiments are held in following time :

6:30–7:30

11:30–12:30

14:30–15:30

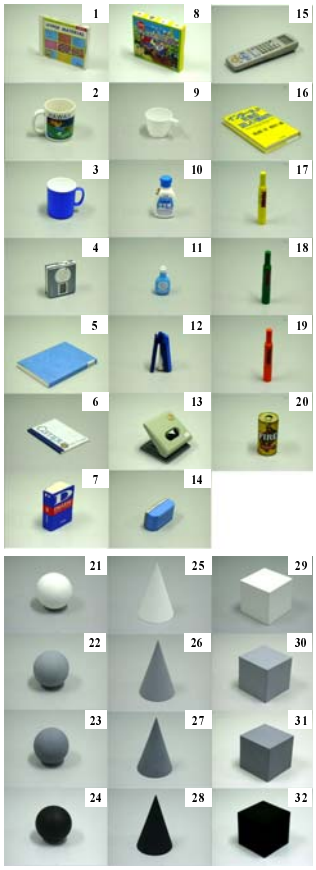

Figure 4.11. Target objects in experiments

I used 33 objects (illustrated in figure 4.11), consisting of 20 utility objects (1 – 21) and 12 standard objects(22 – 33). Standard objects consist of spheres(22 – 25), corns(26 – 29) and cubes(30 – 33). Utility objects were arbitrary samples with follwing policies. 1) A set of utility objects was made to include wide variety of features (surface material, color pattern, and shape) so as to avoid a specific result. 2) Maximal diagonal length was constrained from 5 cm to 25 cm so as to be captured fully in the frame of cameras and not to be captured in too small a size. Standard objects in each shape group are colored in three gradient steps (30%, 60%, and 100%) of black.

Each object was set on an object table at a distance of 35cm from the installation location of each camera. *ObjectCam2* and *ObjectCam* captured the object to create a $NearbyMask$ of the object image. The $NearbyMask$ of *ObjectCam2*, covering only a region of the nearby object image, was used in the masking phase in figure 4.9. The

$NearbyMask$ of *ObjectCam* was created by simply binarizing the IR image *ObjectCam* itself captures. Each threshold for binarizing was adjusted to an appropriate value in each environment based on heuristics.

To evaluate the quality of the $NearbyMask$ in comparing *ObjectCam2* and *ObjectCam*, I have calculated a noise ratio and the missing ratio of each $NearbyMask$. The area of noise S_n and that of the missing S_m indicate the quality of the $NearbyMask$. These values are calculated using a $CorrectMask$, which is a template for evaluating the $NearbyMask$. I manually create the $CorrectMask$ for each $NearbyMask$ from the source image of the object. S_n is defined as a partial area value of the $NearbyMask$ covered by the background area of the $CorrectMask$. S_m is defined as the partial area value of the $CorrectMask$ excluding the area covered by the $NearbyMask$. When S_c is a common area value between $CorrectMask$ and $NearbyMask$, the noise ratio P_n and the missing ratio P_m of the $NearbyMask$ are denoted as below :

$$P_n = \frac{S_n}{S_c}. \tag{4.24}$$

$$P_m = \frac{S_m}{S_c + S_m}. \tag{4.25}$$

The value of P_n in case of $S_c = 0$ is defined as "No value" that indicates there is no extracted area in the correct region of the nearby object.

Results

Figure 4.13 denote comparisons of noise ratio between ObjectCam and ObjectCam2 in experimental environments (a) – (d). Each marker on Figure 4.13 denotes experimental result of each object in a certain experimental condition. The horizontal axis denotes the noise ratio of ObjectCam, and the vertical axis denotes the noise ratio of ObjectCam2. Both axes in Figure 4.13 (a) (c) are logarithmic axes. Additional lines (L1) displays locations that the noise ratio by ObjectCam2 is one-tenth of that by ObjectCam. (L2) denotes locations the the noise ratio by ObjectCam2 is ten times bigger than that by ObjectCam.

From experimental results, I found that ObjectCam2 has created more accurate NearbyMask than that created by ObjectCam in experimental environment (a) and (c). On the other hand, Both cameras have achieved under 10% of noise ratio in experimental environment (b) and (d), as denoted in Figure 4.13 (b) and (d).

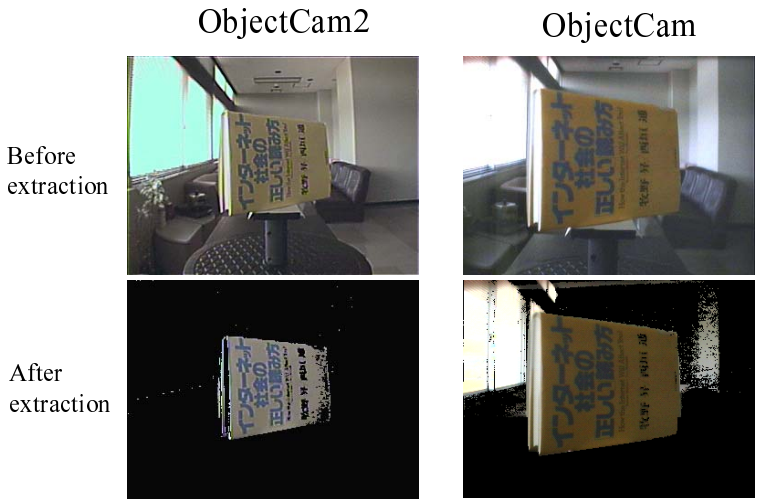

Figure 4.12. Experimental result
(*NearbyColor* of object(16))

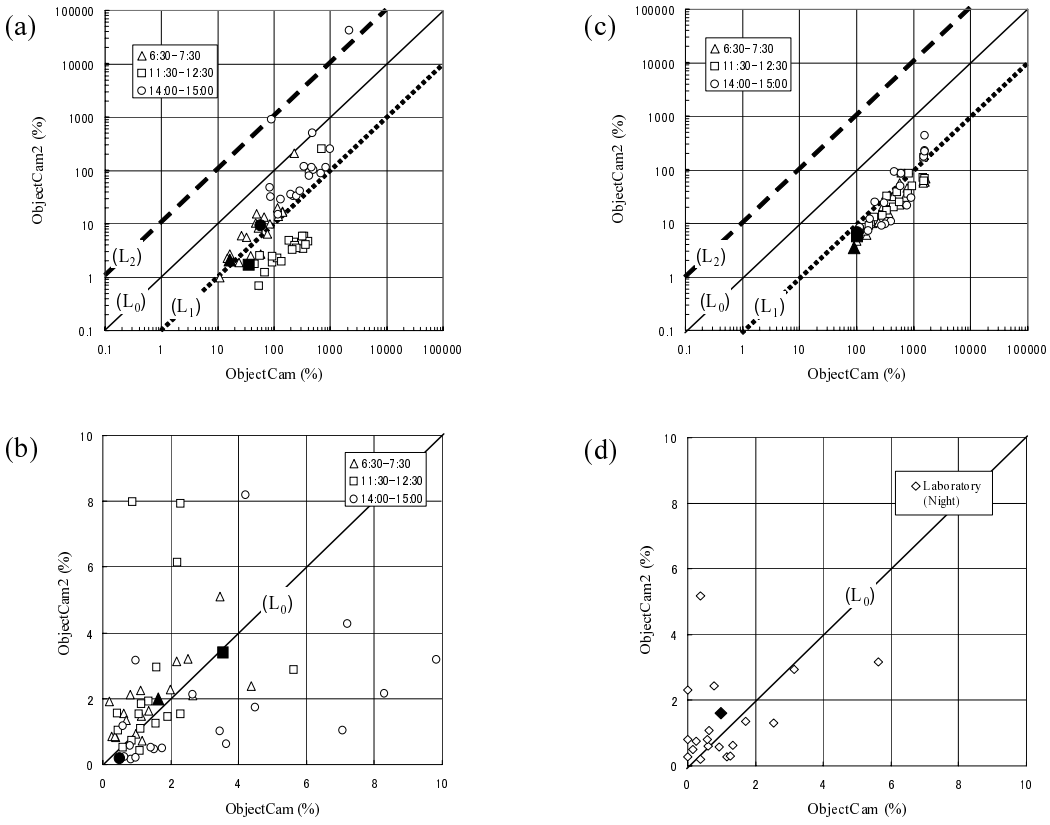

Figure 4.13. Comparison of extraction accuracy

Following results are not included in Figure 4.13 :

[No value] Object(9) in environment(c) at time(14:00 – 15:00)

In all experimental environment, each NearbyMask of utility object(16) extracted by ObjectCam2 has noise ratio lower than 10%, as black markers plotted on Figure 4.13. Figure 4.12 denotes images of object(16) extracted by ObjectCam and ObjectCam2. In the result of using ObjectCam, not only nearby regions but background regions illuminated by sunlight have been also extracted. In contrast, the result of using ObjectCam2 has only nearby regions without subtracted background regions.

4.3.3 Discussions

The noise ratio of the $NearbyMask$ must be decreased to achieve a robust identification of a trigger object using *ObjectCam2*. From equation (4.24), the noise ratio should be decreased by increasing S_c. Increasing S_c means decreasing the missing area at the same time.

I found a problem such that the dark colored surfaces of object images were dropped out in the outdoor environment. The problem has been indicated in experimental results of gradient black colored standard objects and cyan colored standard objects. Results in utility object (18) are also caused by similar problem. One of the ways to respond to the problem is to enhance the luminance of active IR LEDs to capture a reflected active IR image even if the surface reflects only a little IR light. To avoid extracting a background region illuminated by strong active IR light, the luminance of each pixel in the object image has to be corrected dynamically into a proper value by considering the color-specific reflectivity of IR light based on the color information of $NIRcolor$.

I found another problem with object (9) that only retains a half-transparent white surface. The object image of object (9) dropped out almost completely in the outdoor experiment of *ObjectCam2*. Strong sunlight was transmitted from the background of object (9), which caused a partial saturation condition of $NIRcolor$ in the region of the object. If $NIRcolor$ has been saturated partially, the luminance of the saturated area in DI will tend to be turned down. This "partial saturation condition" problem is common to the partial reflection of environmental light on a part of the object surface, and is also common to the sky region captured in an object image. To solve the partial

saturation problem completely, not only $\delta_{IRcolor}$, but also $\delta_{NIRcolor}$ in equation (4.23), has to be constrained so that $NIRcolor$ does not have a saturated area.

A dilemma arose, however, in solving the darkcolored object extraction problem and the partial saturation problem simultaneously. The dark colored object extraction problem needs bright $NIRcolor$ to utilize the color information of $NIRcolor$. On the other hand, the constraint of $\delta_{NIRcolor}$ shows that $NIRcolor$ contains less color features than under the normal condition. To improve the robustness of *ObjectCam2* in extracting a nearby object image under sunlight illumination, the dilemma must be solved.

Chapter 5

Concluding Remarks

This chapter mentions total conclusions of this thesis from results illustrated in above chapters. Experimental results have shown that the system design of *I'm Here!* is effective and implement-able to support a user's object-finding task in his/her everyday environment. Future works are proposed to make *I'm Here!* more effective to support everyday object-finding task.

5.1 Contributions

I proposed a system design of *I'm Here!* that contributes to develop a wearable Augmented Memory system to support a user's remembrance of an event where he/she last placed a target object in his/her everyday life. *I'm Here!* is designed to have less difficulty in registration operation of target objects and retrieval operation of a first-person video for the remembrance of the event. The user of *I'm Here!* is no cause for being conscious of the system behaviour constructing an Experience Database that accumulates first-person videos for the remembrance of the event.

Following achievements contribute to develop a wearable Augmented Memory system that can make the user's object-finding task efficient in his/her everyday life :

1. I evaluated requirement for a function of the system to recognize a target object. The accuracy of the object recognition affects the performance of an Experience Database constructed by the system. Experimental results revealed the functional availability of the method to support a user's object-finding task if the system accurately recognizes an object held by him/her. The results also

indicated that *I'm Here!* requires 67% of the accuracy of object recognition.

2. I investigated an appropriate view size of a first-person video so that a user can recognize contexts of the first-person video. The contexts are needed to recall an event that he/she last placed the target object. Experimental results indicated that the appropriate range of view size is from 115 to 125. The number of view size represents the horizontal view angle of the lens recommended for recording of the first-person video.

3. I have developed an *ObjectCam2* camera device, which is worn by a user, to extract a nearby object image from its background image. An object-triggered memory augmentation system can simplify its system construction by using *ObjectCam2* for both video capturing and object identifying. Integrating the color/IR image-capturing function and the blinking function of IR LEDs, the *ObjectCam2* performs well in extracting an object image held by a user in his/her everyday environment. The saturation problem in extracting the object image with a highlighted surface has been cleared by the exposure time controlling function of *ObjectCam2*.

5.2 Future works

This thesis only describes methods to support a user searching for a target objet placed by him/her. If the target object has been shared by a number of persons, each person tends to meet with "a multi-person object-finding task" that is a task to find a target object last moved and placed by another person. It is a future work to develop *Multi-user I'm Here!* supporting the multi-person object-finding task. *Multi-user I'm Here!* should have function to manage a number of Experience Databases each of them records experiences of each user.

Following problems have to be solved to design *Multi-user I'm Here!* :

Registration of target objects : Information of target objects should be shared between a number of *Multi-user I'm Here!* so that every Experience Databases will be constructed with common names of the target objects. It is necessary to discuss a rule of object registration for a number of users. A method to share a common Object Database is required.

Identification of a user who last moved the target object : If *Multi-user I'm Here!* searches a user who last moved the target object from all users, the system should check huge number of Experience Databases. It is necessary for real-time processing of *Multi-user I'm Here!* to develop a method to reduce the number of candidate users.

Respect for each user's privacy : *Multi-user I'm Here!* should be designed to respect a user's privacy when the system utilizes his/her Experience Database to solve another user's object-finding task. The system also has to respect the privacy of a user searching for a target object if he/she doesn't want to be known to others that he/she is searching for the object.

References

Bergener, T. and Dahm, P. (1997). A framework for dynamic man-machine interaction implemented on an autonomous mobile robot. *Proceedings of ISIE'97 IEEE International Symposium on Industrial Electronics.*

Brewer, W.F. and Treyens, J.C. (1981). Role of Schemata in memory for places. *Cognitive Phychology*, Vol. 13, pp.207–230.

Davenport, L. (2001). ORDER FROM CHAOS : A 6-Step Plan for Organizing Yourself, Your Office, and Your Life. Three Rivers Press, New York, 2001.

Gemmell, J., Williams, L., Wood, K., Bell, G. and Lueder, R. (2004). Passive Capture and Ensuing Issues for a Personal Lifetime Store. *Proceedings of The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)*, pp. 48–55.

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

Jebara, T., Schiele, B., Oliver, N. and Pentland, A. (1998). *DyPERS*: Dynamic Personal Enhanced Reality System. *MIT Media Laboratory, Perceptual Computing Technical Report #463.*

Kawashima, T., Nagasaki, T. and Toda, M. (2002). Information Summary Mechanism for Episode Recording to Support Human Activity. *Proceedings of the International Workshop on Pattern Recognition and Understanding for Visual Information Media*, pp.49–56.

Kawamura, T., Kono, Y. and Kidode, M. (2001) A Novel Video Retireval Method to Support a User's Recollection of Past Events Aiming for Wearable Information Playing. *Proceedings of the 2nd IEEE Pacific-Rim Conference on Multimedia*, pp.24–32.

Kawamura, T., Kono, Y. and Kidode, M. (2003a). *Nice2CU*: Managing a Person's Augmented Memory. *Proceedings of 7th IEEE International Symposium on Wearable Computers (ISWC2003)*, pp.242–243.

Kawamura, T., Fukuhara, T., Takeda, H., Kono, Y. and Kidode, M. (2003b). Ubiquitous Memories: Wearable Interface for Computational Augmentation of Human

Memory based on Real World Object. *Proceedings of 4th International Conference on Cognitive Science (ICCS2003)*, pp.273–278.

Kidode, M. (2002). Design and Implementation of Wearable Information Playing Station. *Proc. 1st CREST Workshop on Advanced Computing and Communicating Techniques for Wearable Information Playing*, pp.1–5.

Kono, Y., Kawamura, T., Ueoka, T., Murata, S. and Kidode, M. (2003). *SARA*: A Framework for Augmented Memory Albuming Systems. *Proceedings of the 2nd CREST Whorkshop on Advanced Computing and Communicating Techniques for Wearable Information Playing*, pp.20–34.

Lamming, M. and Flynn, M. (1994). Forget-me-not: Intimate Computing in Support of Human Memory. *Proc. FRIENDS21: International Symposium n Next Generation Human Interface*, pp.125–128.

Leibe, B. and Schiele, B. (2003). Analyzing Appearance and Contour Based Methods for Object Categorization. *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, Wisconsin.

Mann, S. (1997). Wearable Computing: A First Step Toward Personal Imaging. *Computer*, Vol. 30(2), pp.23–30.

Nayar, S. K., Nene, S. A. and Murase, H. (1996). Real-Time 100 Object Recognition System, *Proceedings of the ARPA Image Understanding Workshop*, pp. 22–28.

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

Nickerson, R.S. and Adams, M.J. (1979). Long-term Memory for a Common Object. *Cognitive Psychology*, Vol. 11, pp.283–307.

Rhodes, B. (1997). The Wearable Remembrance Agent: a System for Augmented Memory. *Proceedings of the 1st International Symposium on Wearable Computers (ISWC'97)*, pp.123–128.

Schiele, B. and Crowely, j. l. (1996). Probabilistic Object Recognition using Multidimensional Receptive Field Histograms, *Proceedings of the 13th Conference on Pattern Recognition*, Vol. B, pp. 50–54.

Shingaki, N., Nojima, H., Kitabata, M., Onozawa, A. and Sato, K. (2003). Analysis of human-object interaction process in ubiquitous environment. *Proceedings of IPSJ SIG Technical Reports*, Vol.2003 No.115, pp.183–188. (in Japanese)

Shinnishi, M., Iga, S. and Higuchi, F. (1999). Hide and Seek: Physical Real Artifacts which Responds to the User. *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics*, Vol. 4, pp. 84–88.

Sumi, Y,, Ito, S., Matsuguchi, T., Fels, S. and Mase, K. (2004). Colllaborative Capturing and Interpretation of Interactions. *Proceedings of Pervasive 2004 Workshop on Memory and Sharing of Experiences*, pp. 1–7.

Terai, H. and Miwa, K. (2004). Analysis of Search Process for Missing Belongings Based on Studies on Insight Problem Solving. *Cognitive Studies*, Vol. 11(3), pp.262–269. (in Japanese).

Toda, M., Nagasaki, T., Iijima, T. and Kawashima, T. (2003). Structural Representation of Personal Events. *Proceedings of the ISPRS International Workshop on Visualization and Animation of Reality-based 3D Models.*

Tulving, E. and Thomson, D.M. (1973). Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological Review*, Vol. 80, pp.352–373.

Ueoka, R., Hirota, K. and Hirose, M. (2001). Wearable Computer for Experience Recording. *Proceedings of the 11th International Conference on Artificial Reality and Teleexistence.*

Weiser, M. (1991). The Computer for the 21th. *Scientific American*, 265(3), pp.94–104.

Winograd, E. and Soloway, R. M. (1986). On Forgetting the Locations of Things Stored in Special Places. *Journal of Experimental Psychology: General*, Vol. 115, pp. 366–372.

Wren, C. R., Azarbayejani, A., Darrel, T. and Pentland, A. (1997). Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 780–785.

Publications

Journals

1. **Takahiro Ueoka**, Tatsuyuki Kawamura, Yasuyuki Kono, Masatsugu Kidode: *I'm Here!* : Wearable Interface System for Improving Efficiency of Finding Objects, *Journal of Human Interface Society*, ISSN 1344–7262, Vol. 6, No. 3, pp. 275–285, August 2004, (in Japanese).

Book

1. Tatsuyuki Kawamura, **Takahiro Ueoka**, Yasuyuki Kono, Masatsugu Kidode : Wearable and Ubiquitous Video Data Management for Computational Augmentation of Human Memory. *In Sagarmay Deb (ed) Video Data Management and Information Retrieval*, IRM Press, pp.33-76, ISBN1-59140-571-8, 2004.

International Conferences

1. **Takahiro Ueoka**, Tatsuyuki Kawamura, Shigeyuki Baba, Shinichi Yoshimura, Yasuyuki Kono, Masatsugu Kidode: Wearable Camera Device for Supporting Object-Triggered Memory Augmentation, *Proceedings of the 3rd CREST/ISWC Workshop on Advanced Computing and Communicating Techniques for Wearable Information Playing*, pp.46–53, Arlington, Virginia, USA, October 31, 2004.

2. **Takahiro Ueoka**, Tatsuyuki Kawamura, Yasuyuki Kono, Masatsugu Kidode: Functional Evaluation of a Wearable Object Remembrance Support System, *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME2004)*, Taipei, Taiwan, June 27–30, 2004.

3. Tatsuyuki Kawamura, **Takahiro Ueoka**, Yasuyuki Kono, Masatsugu Kidode: Relational Analysis among Experiences and Real World Objects in the Ubiquitous Memories Environment. *Proceedings of the Pervasive 2004 Workshop on Memory and Sharing of Experiences*, pp.79-85, April 20, 2004.

4. Yasuyuki Kono, Tatsuyuki Kawamura, **Takahiro Ueoka**, Satoshi Murata, Masatsugu Kidode: Real World Objects as Media for Augmenting Human Memory, *Proceedings of the Workshop on Multi-User and Ubiquitous User Interfaces (MU3I)*, pp.37–42, Funchal, Madeira Island, Portugal, January 13, 2004.

5. **Takahiro Ueoka**, Tatsuyuki Kawamura, Yasuyuki Kono, Masatsugu Kidode: *I'm Here!*: a Wearable Object Remembrance Support System, *Proceedings of the 5th International Symposium on Human Computer Interaction with Mobile Devices and Services (MobileHCI03)*, pp.422–427, Udine, Italy, September 8–11 2003.

6. Yasuyuki Kono, Tatsuyuki Kawamura, **Takahiro Ueoka**, Satoshi Murata, Masatsugu Kidode: *SARA*: A Freamwork for Augmented Albuming Systems, *Proceedings of 2nd CREST Workshop on Advanced Computing and Communicating Techniques for Wearable Information Playing*, pp.30–34, Nara, Japan, May 23–24, 2003.

Domestic Conferences (in Japanese)

1. Tatsuyuki Kawamura, **Takahiro Ueoka**, Yutaka Kiuchi, Yasuyuki Kono, Masatsugu Kidode: Probabilistic Hue Histogram Integrated with Saturation and Value, *Proceedings of the 15th IPSJ-SIG-VIR*, pp.49–52, Osaka, Japan, October 31, 2003.

2. **Takahiro Ueoka**, Tatsuyuki Kawamura, Yasuyuki Kono, Masatsugu Kidode: *I'm Here!* - A Wearable Object Remembering Support System, *Proceedings of the 65th Annual Conference of IPSJ 2003*, Vol.5, pp.179–182, Tokyo, Japan, March 25–27, 2003.

3. Yasuyuki Kono, Tatsuyuki Kawamura, **Takahiro Ueoka**, Satochi Murata, Norimichi Ukita, Masatsugu Kidode: Towards the Wearable Video Diary - Memory Retrieval, Editing, Transportation , and Exchange in Daily Life, *Proceedings of the IEICE PRMU2002-178*, pp.55-60, Nara, Japan, January 16–17, 2003.

4. Tatsuyuki Kawamura, **Takahiro Ueoka**, Norimichi Ukita, Yasuyuki Kono, Masatsugu Kidode: Toward a Memory Aid System in the Wearable Information Playgin Project, *Proceedings of the 3rd Meeting for Youth Community 2002 (MYCOM2002)*, Shiga, Japan, June 17–18, 2002.

5. Tatsuyuki Kawamura, Hideaki Takeda, Kazunori Terada, Tomohiro Fukuhara, Masaki Chikama, Kingo Koshiishi, **Takahiro Ueoka**, Masahiro Hamasaki: AgentBox: a Hardware for a Novel Interaction between Human and Artifacts, *Proceedings of the 16th Annual Conference of JSAI*, Tokyo, Japan, May 28–31, 2002.

6. **Takahiro Ueoka**, Tatsuyuki Kawamura, Norimichi Ukita, Yasuyuki Kono, Masatsugu Kidode: An Object Remembrance Support System using a Wearable Camera System, *Proceedings of Interaction 2002*, pp.63–64, Tokyo, Japan, March 6–7, 2002.

7. **Takahiro Ueoka**, Tatsuyuki Kawamura, Yasuyuki Kono, Masatsugu Kidode: Basic Study for a Wearable Object Registration and Retrieval System, *Proceedings of ITE Technical Report*, Vol.26, No.7, pp.25–30, Tokyo, Japan, January 25, 2002.

8. Akihiro Terabe, **Takahiro Ueoka**, Tatsuyuki Kawamura, Yasuyuki Kono, Masatsugu Kidode: A New Input Interface for Wearable Computing, *Proceeding of the 15th Annual Conference of JSAI*, Shimane, Japan, May 22–25, 2001.

Technical Reports

1. **Takahiro Ueoka**, Tatsuyuki Kawamura, Norimichi Ukita, Yasuyuki Kono, Masatsugu Kidode: Remembering the Object Location with a Wearable Vision Interface, *Information Science Technical Report #NAIST-IS-TR2002017*, Nara Institute of Science and Technology, 2002.

Patent (in Japanese)

1. Y. Kono, M. Kidode, **T. Ueoka**, T. Kawamura: Close region image extraction device and close region image extraction method, ヨーロッパ特許(EPC), 公開番号 EP 1 465 415 A1, 平成15年12月17日出願, 2003（国内優先 特開2004-304718, 平成15年4月1日出願, 2003）.

2. Y. Kono, M. Kidode, **T. Ueoka**, T. Kawamura: Close region image extraction device and close region image extraction method, アメリカ合衆国特許, 出願番号 10/731982, 平成15年12月10日出願, 2003（国内優先 特開2004-304718, 平成15年4月1日出願, 2003）.

3. 河野恭之, 木戸出正継, **上岡隆宏**, 河村竜幸: 近接領域画像抽出装置及び近接領域画像抽出方法, 日本国特許, 特開2004-304718, 平成15年4月1日出願, 2003.