

Supporting On-Demand Experience Segmentation in the Ubiquitous Memories Environment

Satoshi Murata, Tatsuyuki Kawamura, Yasuyuki Kono, and Masatsugu Kidode
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma
Nara, 630-0192, JAPAN
E-mail {satosh-m,kawamura,kono,kidode}@is.naist.jp

Abstract

This paper newly proposes a wearable system in the Ubiquitous Memories (UM) environment we have already developed, where users can enclose/disclose their experiences in/from related objects. We have designed an intelligent wearable system to support on-demand experience segmentation so that an object can be enclosed by the following two processes 1) continuously recording the user's viewpoint images, and 2) detecting the starting point of an experience the user wants to record just when that experience has ended. We describe design concepts of the proposed system and an experiment to confirm its availability to detect the starting point of a certain period of an experience.

1 Introduction

Several kinds of intelligent environments have been proposed, some of which are already implemented; for example, the Robotic Room[1], the Ubiquitous Memories (UM) environment[2], and so on. The Robotic Room is designed to support persons in the room by a set of sensors and cameras. This room monitors persons, stores their behavior information, guide them like as a nurse, and so on. The UM-environment is also implemented where a user directly recognizes his/her experience via a real world object related to its experience. In the UM-environment, a user wears a wearable system that continuously captures the user's viewpoint video. Here, we term this kind of video, which includes the contexts of the user's experiences, an "experience video."

These intelligent environments can get too much data to support persons. Therefore an intelligent data management is required in order to improve their usefulness. We propose an intelligent wearable system which can segment a user's experience video on-demand. The purpose of this system is to automatically segment a scene of experience video. The problem for realizing the system is how the length of a

video is determined depending on the context of the experience itself. Here, we propose a method to detect the starting point of the experience the user wants to record. When the experience the user wants to record has ended, he/she decide the ending point of the experience by pushing a button. Then the system supports the experience segmentation by detecting the starting point. We employ multiple sensors and the method, which is to match the time-series data read from multiple sensors, to detect the starting point.

In the section 2, we describe our newly developed the UM-environment with primitive functions. And we propose an on-demand experience segmentation system in the section 3. Finally we demonstrate experimental results with some discussions.

2 UM-Environment

We have proposed a conceptual design for ideally and naturally bridging the space between experience video and human memory by regarding each real world object as part of an experience video archive. To seamlessly integrate between human experience and experience video, we believe that providing users with natural actions for storing/retrieving experience video is important. A "human hand" plays an important role in integrating the experience video into objects. The human body is used as media for both perceiving the current context (event) as a memory and for propagating the memory to an object, i.e., the memory travels all over the user's body like electricity and the memory runs out of one of his/her hands in our design. Terms of conceptual actions[3] are defined as follows:

- Enclose
action is shown by two steps of behavior. 1) A person implicitly/explicitly gathers a current context through his/her own body. 2) He/She then arranges contexts as ubiquitous experience video with a real world object using a touching operation.

- Accumulate
denotes a situation where experience videos are enclosed in an object. The situation functionally means that the experience videos are stored in computational storage somewhere on the Internet with links to the object.
- Disclose
action is a reproduction method where a person recalls the context enclosed in an object. The “Disclosure” has a similar meaning of replaying media data.

3 Concept of On-Demand Segmentation

Experience video segmentation methods employing multiple sensors have been studied[4][5][6][7]; e.g., Global Positioning Systems (GPS), gyro sensors, acceleration sensors, brainwave sensors, thermometers, heart beat sensors, and so on. Most of these are based on well-known principles; e.g., such as “alpha-blocking” (the alpha waves decline) which is observed when a user is excited, “RR-interval variability” (the interval of the peaks of the electrocardiogram) which decrease when a user feels stressed, etc.

In our experiments so far, however, the user has not explicitly shown his/her intentions when requiring a segmentation of an experience video. Therefore the aim of this study is to realize a system that is able to dynamically generate implicit rules via the user’s intentions.

3.1 Focus of this Research

This paper focuses on the type of scenes where a user wants to record his/her experiences just after they occur. The model of the scene is shown in Figure 1.

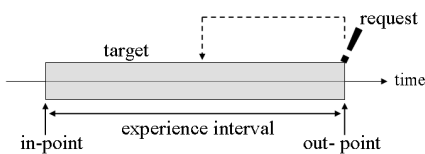


Figure 1: Focus of research in experience segmentation

The “request” is the desire of a user to record his/her experience. The “target” is the experience he/she wants to record. The “experience interval” is the period during the target. The “in-point”/“out-point” is the position at which the target starts/ends.

The aim of this research is to realize a system which retrieves an in-point when a request arises (= an out-point).

3.2 Architecture of the Proposed Wearable System

The aim of this research is to realize a system, which retrieves an in-point when a user decides an out-point and adjusts the criterion for retrieving when a user modifies the in-point. Therefore the system for on-demand segmentation of an experience video must have the following three capabilities:

- Retrieving the in-point, when the user decides the out-point
- Adjusting the in-point, when the user modifies it
- Adjusting the criterion for retrieving the modified in-point

We employ a method for matching the time-series data read from multiple sensors in order to retrieve the in-point because we assume that every time-series data read from multiple sensors in experience intervals implies features caused by the user’s own cognition. Based on this assumption, the interval of the data, which is most similar to the other data around the in-point, is retrieved, and the central position of the interval is estimated at the in-point.

We employ adjustable processes for converting the time-series data of multiple sensors, and the adjustable groups of the data, in which the data use the same process for estimating the in-point. When a user modifies the in-point, the processes or the groups are adjusted in order to correspond to the modified in-point.

We propose the following system with the required capabilities for on-demand segmentation of experience video (Figure 2).

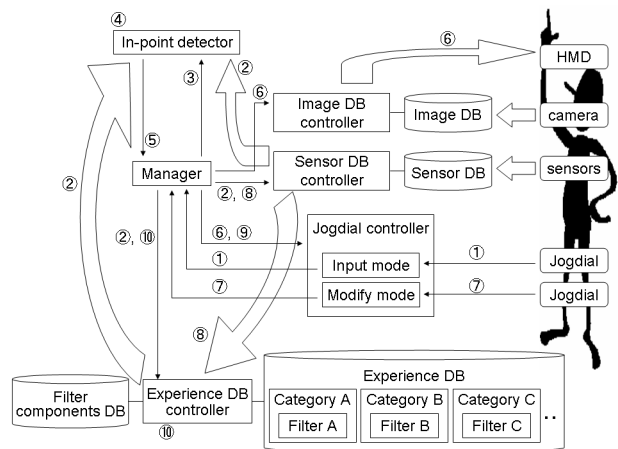


Figure 2: System block diagram

The system is composed of a “manager,” an “in-point detector,” a “jogdial controller,” a “sensor

database,” an “image database,” and an “experience database” with certain “categories.” A sensor database, an image database, and an experience database each have an individual “controller.” An experience database controller has a “filter components database.” Each category has a “filter” and a “typical data.” The manager sends commands to other parts of the system. The in-point detector estimates the in-point by matching the data. A filter is composed of various filter components (differential, integral, etc.) and is an adjustable process for converting the data. A category is an adjustable group of the data. The experience database stores the segmented time-series data of multiple sensors. The sensor database and the image database continuously record the time-series data of the multiple sensors and the user’s viewpoint images respectively. The jogdial controller gets commands from the jogdial. The sensor/image database controller sends the data stored in the sensor/image database. The experience database controller gets the segmented time-series data of multiple sensors, sends the filters and the typical data which categories have, and builds filters and categories in order to correspond to the modified in-point.

The required process of the system is completed by the following ten steps.

- ① A user pushes a jogdial as an out-point when he/she wants to record his/her experience. The jogdial controller sends the signal with the time of the out-point after detecting the in-point.
- ② The manager gives an order to the sensor database controller and the experience database controller. Next, the sensor database controller sends the sensor data, which is recorded before the time of the out-point, to the in-point detector. The experience database controller sends the filters and the typical data, which categories have, to the in-point detector.
- ③ The manager commands the in-point detector to start the estimative process.
- ④ Each filter is applied to the data sent from the sensor database controller, and the data are matched to each typical data. The most similar position decided from the result of matching is estimated as the in-point.
- ⑤ The in-point detector sends the time of the estimated in-point back to the manager.
- ⑥ The command of the manager enables the image database controller to replay the video from the

estimated in-point to the out-point. The manager makes the jogdial controller switch the input mode to the modify mode.

- ⑦ The user modifies the estimated in-point to the position he/she wants to segment by using the jogdial. The jogdial controller detects the modification signal, and sends the signal with the time of the modified in-point to the manager.
- ⑧ The manager requires the sensor data controller to send the sensor data to the database controller. Next, the sensor database controller sends the sensor data, which is recorded between the modified in-point and the out-point, to the experience database controller.
- ⑨ The manager makes the jogdial controller switch the modify mode to the input mode.
- ⑩ The manager commands the database controller to start the rebuilding process. Next, the database controller rebuilds some of the filters or categories in order to adjust to the modified in-point.

The rebuilding process allows a more exact effect to be acquired by the modification of fewer filters or categories. Therefore, the rebuilding process is expanded from a filter to a category.

In this paper, “to be able to be applied” means that the estimated error of all data belongs to a category falling below the threshold. Furthermore, a category must have two or more data, because the method to match the data for estimating an in-point needs at least two data.

The rebuilding process is composed of the following five steps (Figure 3):

- A. If there are not any registered data, no category is built. If there is only one piece of registered data, only one category with a filter is built.
- B. If there are any filters (i.e. there are more than two registered data), the filter which can be applied to the new data is searched for.
- C. **(Modify filter function)** The category with the minimum estimated error is searched for, and the filter belonging to the category is modified in order to be an applicable filter.
- D. **(Rebuild filter function)** While there are candidates of the typical data in the category, a filter is rebuilt with changed typical data.
- E. **(Unify category function)** After a category is divided, a unifying category is tried. The candidate is the category with the maximum estimated error.

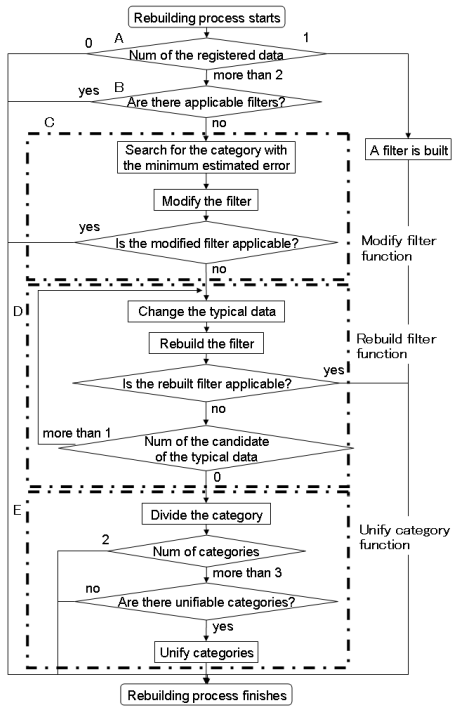


Figure 3: Rebuilding process flowchart

4 Experimental Results

The purpose of the experiment is to investigate the possibility that the in-point can be retrieved by matching the time-series data of brainwave sensors and electrooculogram (EOG) sensors.

4.1 Analytic Methods

In this paper, we employ the Dynamic Time Warping (DTW) method[8] to match time-series data. The DTW method computes the distance between two time-series data with non-linear time normalization.

In this paper, we employ the power, the integral, and the differential as filter components.

- Power

The power $P_w(t)$ of the time-series data $F(T)$ at point $T = t$ is calculated by

$$P_w(t) = \sqrt{\sum_{T=t-w}^t F(T)^2} .$$

- Integral

The integral $I_w(t)$ of the time-series data $F(T)$ at point $T = t$ is calculated by

$$I_w(t) = \sum_{T=t-w}^t F(T) .$$

- Differential

The differential $D_w(t)$ of the time-series data $F(T)$ at point $T = t$ is calculated by

$$D_w(t) = F(t) - \sum_{T=t-w}^{t-1} F(T)/w .$$

In each equation, the value of w is a parameter of each filter component. In this research, we use these filter components in a cascade.

The analytic process is shown in Figure 4.

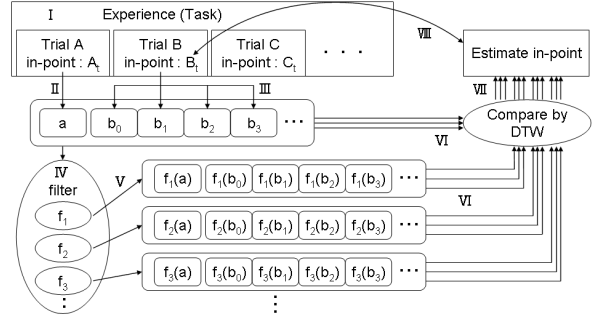


Figure 4: Analytic process

- I. In this experiment, a subject has some trials in each task. We regard these trials as the same kind of experiences. Each data set read from multiple sensors in each task has a right in-point which is decided by the subject itself.
- II. The data set “a” is the time-series data for T_{len}^1 , which is separated from the data read from multiple sensors in trial A so that the center of the length is the right in-point A_t .
- III. The data set “ b_0 ” is the time-series data for T_{len} , which is separated from the data read from multiple sensors in trial B so that the center of the length is the right in-point B_t . Furthermore the data “ b_1 ,” “ b_2 ,” “ b_3 ,” ... are the time-series data for T_{len} , which are separated so that the centers of the length are shifted for t_1 milliseconds, t_2 milliseconds, t_3 milliseconds, ... from the right in-point B_t .
- IV. The filters “ f_1 ,” “ f_2 ,” “ f_3 ,” ... whose parameters differ individually are prepared.
- V. The filters f_1, f_2, f_3, \dots are individually applied to the data set $a, b_0, b_1, b_2, b_3, \dots$.

¹In this paper, T_{len} is 2 seconds.

- VI.** In each channel, the following procedure is implemented for every applied filter;
 a is compared with $b_0, b_1, b_2, b_3, \dots$,
 $f_1(a)$ is compared with $f_1(b_0), f_1(b_1), f_1(b_2), f_1(b_3), \dots$,
 $f_2(a)$ is compared with $f_2(b_0), f_2(b_1), f_2(b_2), f_2(b_3), \dots$,
 $f_3(a)$ is compared with $f_3(b_0), f_3(b_1), f_3(b_2), f_3(b_3), \dots$.
- VII.** The data set which has the minimum result of comparison by DTW is tagged b_α , and the point which is the center of the length of b_α is decided as the estimated in-point B_α .
- VIII.** The estimated error is calculated by comparing the right in-point B_t with the estimated in-point B_α .
- IX.** The process between II. and VIII. is repeated in all combinations of trials.

4.2 Experiments

We conduct an experiment with three tasks. The experiment employs brainwave and electrooculogram sensors for events with mind movement. A man is selected as a test subject in the experiment.

We record the time-series data of an electroencephalogram (EEG) and an electrooculogram with 100 Hz. We use a 10-20 electrode system (Figure 5), and select 7 measuring points to monitor the electroencephalogram signals. The effect on vision is seen from the electroencephalogram signals at O_1 and O_2 . The effect on hearing is seen from the electroencephalogram signals at T_3 and T_4 . The effect on emotion and decision is seen from the electroencephalogram signals at F_{p1}, F_{p2} , and F_8 [9]. The electroencephalogram data is divided into alpha and beta waves before the analysis.

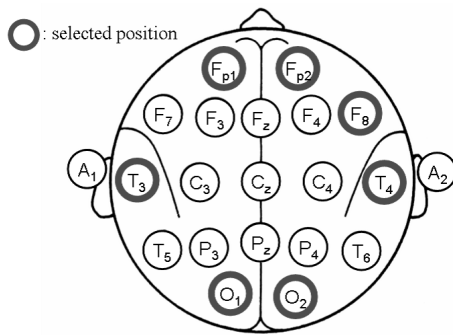


Figure 5: 10-20 electrode system

Before the experiment, the camera recording the subject’s viewpoint images, the brainwave and elec-

trooculogram sensors, and the button recording his requests are attached to him. During the experiment, the out-point is set when the subject pushes the button. After the experiment, the subject sets the in-point by referring to his viewpoint images. The analysis target is the data of multiple sensors from the position 10 seconds before from the in-point to the out-point.

This experiment includes the following three tasks.

T-1 Watching commercials

The subject watches the video containing commercials. The number of commercials in the video is 62. The in-point/out-point is the position at which the commercial he wants to be reminded of starts/ends.

T-2 Watching CD ranking

The subject watches the video containing the “CD single top 50” and the “CD album top 10.” The in-point/out-point is the position at which the video of the ranking he wants to be reminded of starts/ends.

T-3 Watching a volleyball game

The subject watches the video of a volleyball game (the Japanese team vs. the Brazilian team). The out-point is the position at which the Japanese team scores a point. The in-point is the position at which the rally begins.

4.3 Results

Table 1 shows the analysis result of the beta wave data at O_2 which is selected because the result has higher estimated probabilities than the results of the other sensor channels. In the table, the “FILTER” represents the filter that makes the probability of the in-point estimation highest in each task. “ P_w ,” “ I_w ,” and “ D_w ” respectively express a power, an integration, and a differentiation filter component, and “ w ” expresses a parameter of each filter component. “AVG” and “SD” respectively show the average and the standard deviation of the estimated error. “1sec,” “2sec,” and “3sec” illustrate the estimated probabilities. We assume that a distribution of estimated errors

	<i>FILTER</i>	<i>AVG</i>	<i>SD</i>	1sec	2sec	3sec
$T-1$	–	0.3	7.2	0.1	0.2	0.3
$T-2$	–	–0.3	7.1	0.1	0.2	0.3
$T-3$	–	–2.5	7.2	0.1	0.2	0.3
$T-1$	$D_{10}D_4I_1$	1.3	3.2	0.2	0.4	0.6
$T-2$	$I_1I_1I_1$	–1.2	3.9	0.2	0.4	0.5
$T-3$	$I_1I_1D_{10}$	2.3	3.8	0.2	0.3	0.5

(unit: sec)

Table 1: Result of estimation

is a normal distribution when the estimated probability is calculated. The estimated probability of either “1sec,” “2sec,” or “3sec” is the probability that the absolute value of the estimated error is a value within 1 second, 2 seconds, or 3 seconds.

In Table 1, all estimated probabilities with “FILTER” show a higher quality than those without “FILTER.”

4.4 Discussion

The result of the experiment shows that the “FILTER” makes the probability of the in-point estimation higher although the experiment also shows that the “FILTER” is different in each task. Therefore, we believe that an efficient method for selecting a “FILTER” from a huge number of combinations of filters is necessary.

The O_2 -channel of the electroencephalogram signal has the tendency to make the estimated probability higher than the other channels. The stimulus effects on the vision are seen in the electroencephalogram signal at O_2 . We suppose that the result is caused by the feature of effects on the vision in the experiment, i.e., watching the video. Therefore, a method for selecting the available channels of the multiple sensors from many channels is necessary to make the estimated probability high.

5 Concluding Remarks

In this paper, we proposed an intelligent wearable system to support on-demand experience segmentation so that an object can be enclosed in the UM-environment. We modeled scenario where a user records his/her experience and designed the system with the required capabilities and the available methods. We tested these methods and found them to be successful. The results are presented.

We plan to investigate the following five tasks:

- Analysis of the relationship among channels of multiple sensors
- Addition of the kind of filter components
- Investigation of available sensors
- Design of an efficient method to select the “FILTER”
- Implementation of the proposed architecture

We see the second and fourth tasks as the most important tasks. In this experiment, we employed only three filter components, and we used them in a cascade. The estimated probabilities with “FILTER” show higher quality than those without “FILTER,” but the probabilities of 60% or less is low. Therefore,

filter components need to be added for higher probabilities, and these filter components need to be used in a complicated structure. In addition, an actual application of the proposed system is needed to design an efficient method to select the “FILTER.”

When the proposed system is realized, the system should give the user the segmented video instantaneously, and he/she can enclose the video in the related object in the UM-environment.

Acknowledgement

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).

References

- [1] T. Sato, Y. Nishida, and H. Mizoguchi: “Robotic room: Symbiosis with human through behavior media,” *Robotics and Autonomous Systems*, No. 18, pp. 185-194, 1996.
- [2] T. Kawamura, T. Ueoka, Y. Kono, and M. Kidode: “Relation Analysis among Experiences and Real World Objects in the Ubiquitous Memories Environment,” *Pervasive 2004 Workshop on Memory and Sharing of Experiences*, pp. 79-85, Austria, 2004.
- [3] Y. Kono, T. Kawamura, T. Ueoka, S. Murata, and M. Kidode: “Real World Objects as Media for Augmenting Human Memory,” *Proc. Workshop on Multi-User and Ubiquitous User Interfaces 2004 (MU3I 2004)*, pp.37-42, 2004.
- [4] Y. Sumi, S. Ito, T. Matsuguchi, S. Fels, and K. Mase: “Collaborative Capturing and Interpretation of Interactions,” *Pervasive 2004 Workshop on Memory and Sharing of Experiences*, Austria, 2004.
- [5] K. Aizawa, K. Ishijima, and M. Shiina: “Automatic summarization of wearable video - indexing subjective interest,” *IEEE Pacific Rim Conference on Multimedia (PCM2001)*, China, 2001.
- [6] R. Ueoka, K. Hirota, and M. Hirose: “Wearable Computer for Experience Recording,” *International Conference on Artificial Reality and Telexistence (ICAT2001)*, Japan, 2001.
- [7] J. Healey and R. W. Picard: “StartleCam: A Cybernetic Wearable Camera,” *The Second International Symposium on Wearable Computers*, pp. 42-49, 1998.
- [8] H. Sakoe and S. Chiba: “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” *ICDD Transaction on Acoustics, Speech, and Signal Processing*, Vol. ASSP-26, No. 1, pp. 43-49, 1978.
- [9] H. H. Jasper: “The ten-twenty electrode system of the International Federation,” *Electroencephalography and Clinical Neurophysiology*, 10, 371-375, 1958.