ICAT 2004

Capture and Retrieval of Life-Log

K. Aizawa, S. Kawasaki, T. Ishikawa, T.Yamasaki
University of Tokyo, Dept. of Frontier Informatics and Dept. of E.E.
{aizawa, kawasaki, issy, yamasaki}@hal.t.u-tokyo.ac.jp

Abstract

In wearable computing environments, digitization of personal experiences will be made possible by continuous recordings using a wearable video camera. This could lead to the ``automatic life-log application". However, it is evident that the resulting amount of video content will be enormous. Accordingly, to retrieve and browse desired scenes, a vast quantity of video data must be organized using structural information. We are developing a ``context-based video retrieval system for life-log applications". Our life log system captures video, audio, acceleration sensor, gyro, GPS, annotations, documents, web pages, and emails, and provides functions that make efficient video browsing and retrieval possible by using data from these sensors, some databases and various document data.

Key words: Context, Life log, wearable computer

1. Introduction

Capturing a life log by electronic means enables us to record our daily life in detail. To date we have only relied on our human memory to record and remember our daily experiences. We tend to quickly lose the details of our experiences. We record special occasions such as travels and family events with videos and photos. Excluding such special events, we rarely record the daily experiences that are the major part of our life; what we do at most is write a short diary. Because of developing technologies related to wearable, ubiquitous, mobile multimedia, we believe that we will be able to automatically capture and record our daily experiences. Research on the capture and retrieval of life logs is emerging quickly. Two new workshops [18, 19] focusing on this area were launched in 2004.

We have been investigating capturing a life log by audio and video with various sensors such as a GPS (location), gyros, acceleration sensors (motion), physiological sensors (brain wave), annotations, documents, emails, and web pages. Retrieval is the most important problem for the life log system. The amount of data captured is very large, and the problem is to find or navigate through the data so that we can watch the details of our experiences. Remembering the context of our experiences is easier than remembering details of them, so the context can provide valuable keys for indexing information. We have been developing a context-based retrieval system, by which we can navigate the huge amount of video data by parameters such as time, location, a person's behavior, and through words of annotations, documents, emails and web pages. We can search the video by the names of shops and stores that are contained in the town database.

There have been some works to log a person's life in the area of mobile computing, wearable computing, video retrieval and database. A person's experiences or activities have been captured from many different application points of view.

In one of the earliest works [1], person's activities in a laboratory were recorded. For example, various personal activities such as personal location and encounters with others, file exchange, workstation activities, etc were recorded. The activity record was shown on a user's PDA terminal. Along the same line, Mylifebits project [2] designs a life time store in order to store and manage large variety of data such as documents, scanned memos and papers, visited webs, digitized phone calls, CDs, etc.

In the area of wearable and multimedia computing, continuous recording by a wearable camera have been investigated and additional sensors have been often utilized too [3-16].

As we discuss later, in order to handle enormous amount of data, effective tags are needed. So far, video analysis based features are used in [7, 10]. Audio and additional sensor data are captured and stored for medical nursing [8]. In [5], a wearable diary is developed that even uses RFID tags to recognize objects.

In addition, some physiological sensor data are also utilized so that subjective status of the person can be evaluated to some degree and used for the tags. For example, in [4], switching operation of a wearable camera is controlled by a person's skin conductivity so that scenes when the user is in arousal status are captured. Brainwave is used in our previous work in order to summarize video captured by a wearable camera [13].

Meetings and events have been also captured and analyzed using various sensors [11, 12]. Not only wearable sensors but also sensors of environments are used to analyze the user's interactions.

We focus on continuous capturing our experiences by wearable sensors (camera, microphone, GPS, gyro, acceleration sensors). We also store short annotations, application documents, webs visited. We have been

investigating efficient retrieval based on context made of such various kinds of data [13, 14, 15, 16, 17].

In the following sections, we describe our current capturing system in Section 2 and the functions we realized in the previous studies in Section 3 and integrated use of a town directory for efficient retrieval. In Section 4, spatio-temporal sampling for key frame extraction (summarization) is described. Finally, we conclude the paper and describe our ongoing extensions.

2. Life Log System

Our life log system can capture data from a wearable camera, a microphone, and various sensors that show contexts. The sensors we used are a brain wave analyzer, a GPS receiver, an acceleration sensor, and a gyro sensor. The main unit of the system is a notebook PC. All sensors are attached to the PC through USBs, serial ports, and PCMCIA slots, as shown in Figure 1.

Using the modem, the system can connect into the Internet almost anywhere in Japan via the PHS (Personal Handy-phone System: a versatile cordless/mobile system developed in Japan) network of NTT-DoCoMo. By referring to data on the Internet, the system records “the present weather in the user’s location”, “various news on that day, which were offered by some news sites or some email magazines”, “all web pages (*.html) that the user browsed”, and “all emails that the user transmitted and received”.

The system monitors and controls the following applications, “Microsoft Word”, “Microsoft Excel”, “Microsoft PowerPoint”, and “Adobe Acrobat”. In addition to web browsing and transmission and reception of emails, these applications are the main software that people use on a computer. Because of monitoring and controlling them, when the user opens a document file (*.doc; *.xls; *.ppt; *.pdf) of such applications, the system can order each application to copy the file and store it as text data.

All data are recorded directly by the PC. A software block diagram is shown in Figure 2. Video signals are encoded into MPEG1, and audio signals into MP3 using Direct Show. To simplify the system, we modified the sensors to receive their power from the PC, and customized their device drivers for the life log system.

The user can use their cellular phone as a controller of the operations “start/stop life log recording”. The system receives the user’s operations on his or her cellular phone.

3. Retrieval of Life Log Video

We, human beings, save many experiences as a vast quantity of memories over many years of life while

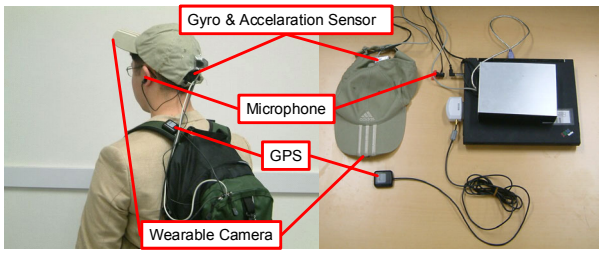

Fig. 1 Wearable devices for life log

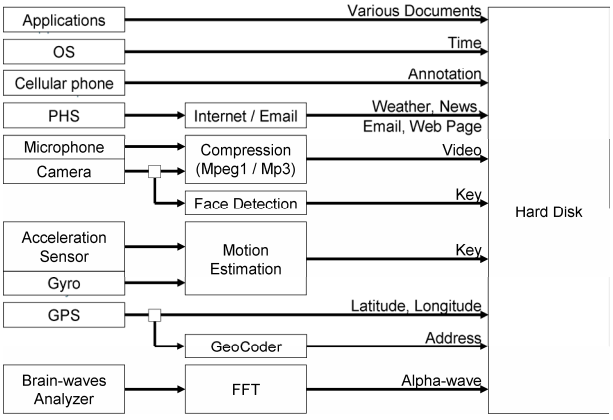

Fig. 2 Block diagram of various types of data

arranging and selecting them, and we can quickly retrieve and utilize the requisite information from our memory. Psychology studies tell that we manage our memories based on contexts. When we want to remember something, we can often use such contexts as keys, and recall the memories by associating them with these keys. For example, to recollect the scenes of a conversation, the typical keys used in the memory recollection process are such context information as ``what, where, with whom, when, how''.

A user may put the following query (Query A). ``On a cloudy day in mid-May when the Lower House general election was held, after making my presentation about life-log, I was called to Shinjuku by the email from Kenji, and I talked with him while walking at a department store in Shinjuku. The conversation was very interesting! I want to see the scene to remember the contents of the conversation.''

In conventional video retrieval the low-level features of image and audio signals of the videos are used as keys for retrieval. Probably, they will not be suitable for queries compatible with the way we query to our memories as in Query A.

However, data from the brain-wave analyzer, the GPS receiver, the acceleration sensor, and the gyro sensor correlate highly with the user's contexts. The life-log system estimates its user's contexts from these sensor data and some database, and uses them as keys for video

retrieval.

Thus, the system retrieves life-log videos by following the way a person recollects experiences from his memories. It is conceivable that by using such context information, the system can produce more accurate retrieval results than by using only audiovisual data. Moreover, each input from these sensors is a one-dimensional signal, and the computational cost for processing them is low.

The interfaces to display the data from the devices are demonstrated in Figure 3. The details of the keys obtained from various devices are provided in the following subsections.

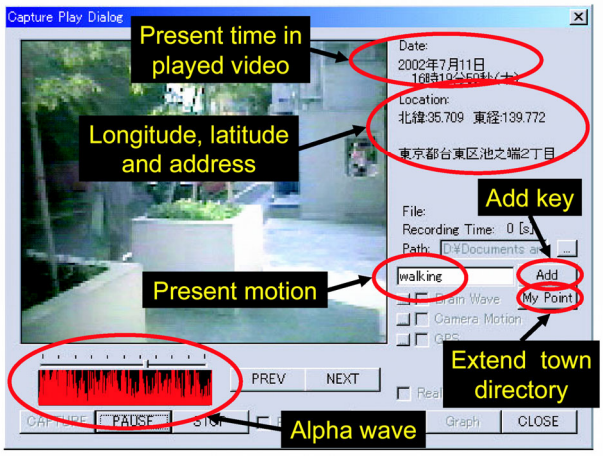

Fig. 3 User interface of capture dialog

3.1 Keys obtained from brain wave analyzer

A brain wave signal, named the □ wave, is acquired from a brain wave analyzer and the data can show the person's arousal status. When a low □ wave or □-blocking is observed, it shows the arousal status of the person, that is, the person interested in something or pays attention. In our previous work [13], very effective retrieval was demonstrated. (However, the brain wave analyzer has not been used often in recent work.)

3.2 Keys obtained from motion sensors

We can capture motion information about our activities from motion sensors including an acceleration sensor and a gyro sensor that are attached to the life log system. These data are used to classify the user motion employing K-Means and HMM algorithms [14]. Hence, from these data, we can identify activities such as walking, running, or not moving.

3.3 Keys obtained from GPS

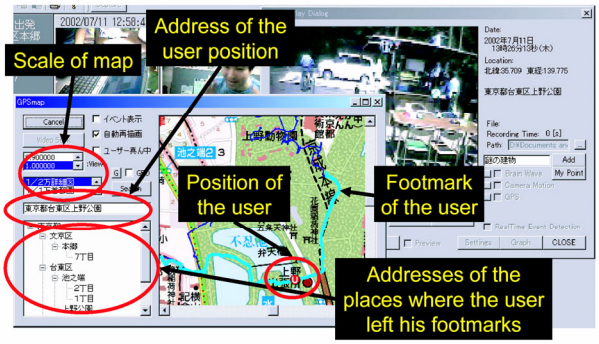

Fig. 4 Location and footprint on map using GPS data

From the GPS signal, the life-log system acquires information about the position of its user as longitude and latitude when capturing a life-log video. The contents of videos and the location information are automatically associated. Longitude and latitude information are one-dimensional numerical data that identify positions on the Earth's surface relative to a datum position. Therefore, they are not intuitively readable for users. The system can convert longitude and latitude into addresses with hierarchical structure using a special database, for example, ``7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan''. The results are information familiar to us, and we use them as keys for video retrieval.

Latitude and longitude information also become information that we can intuitively understand by being plotted on a map as the footprints of the user, and thus become keys for video retrieval. ``What did I do when capturing the life-log video?'' A user may be able to recollect it by seeing his footprints. The system draws the user's footprint in the map under playback using a thick light-blue line, and draws other footprints using thin blue lines on the map. By simply ``dragging his mouse'' on the map, the user can change the area displayed on the map. The user can also order the map to display the other area by clicking arbitrary addresses of all the places where footprints were recorded. The user can watch the desired scenes by choosing arbitrary points of footprints

3.4 Keys obtained from time data

The portable computer also records the time synchronously with the operating system. We can therefore acquire the present time associated with audio/visual information in the life log video. Hence, queries about date and time can be supported by the operating system.

3.5 Data from the Internet

The life log system records data about the weather at the user's location, news on that day, web pages that the user browses, and emails that the user sends and receives while capturing a life log video. These data are

automatically associated with time data. Afterwards, these data can be used as keys for life log video retrieval.

For example, the system informs the user of the weather at the time of recording each video. Thus, the user can choose a video that was captured on a fine, cloudy, rainy or tempestuous day. He or she can choose the day when an arbitrary news event happened, easily and immediately.

Web addresses and emails are stored in the PC. Then, the user can retrieve videos and the content of web pages and emails by keywords that appear in the content.

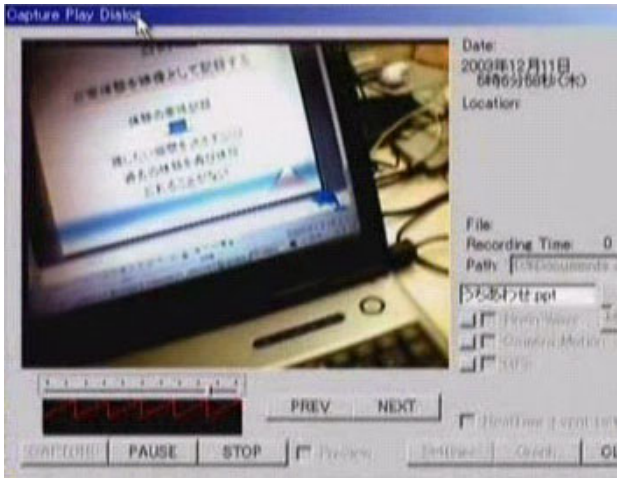

Fig. 5 A result of retrieval from PowerPoint-document

3.6 Data from various applications

All the document files (*.doc; *.xls; *.ppt; *.pdf) that the user opens are copied and saved as text. These copied document files and text data are automatically associated with time data. Afterwards, these text data can be used as keys for life log video retrieval. Alternatively, the system can perform video-based retrieval for such documents including web pages and emails. By clicking the replay of the video of documents, the original document appears.

3.7 Annotations by the user

The user can order the life log system to add retrieval keys (annotation) with an arbitrary name by simple operations on his or her cellular phone while the system is capturing a life log video. The annotations are free-format text. The user operates their cellular phone and easily writes a short message sent to the system via email and the system associates the message to the video based on the time. This function enables the system to identify a scene that the user wants to remember throughout his or her life, and thus the user can easily access the videos that were captured during precious experiences.

4. Combination of Context and a Town Directory

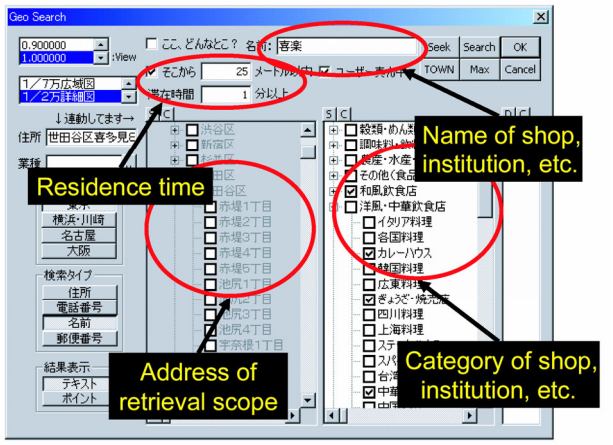

Fig. 6 Retrieval using the town directory

The system is able to use a town directory that enables the user to search video by keywords in the town directory. The database has a vast amount of information about one million or more public institutions, stores, companies, restaurants, and so on in Japan. Except for individual dwellings, the database covers almost all places in Japan including small shops or small companies that individuals manage. In the database, each site has information about its name, its address, its telephone number, and its category with layered structures.

Using this database, a user can retrieve his life-log videos as follows. He can enter the name of a store or an institution, or can input the category. He can also enter the both. For example, we assume that the user wants to review the scene in which he visited the supermarket called ``Shop A", and enters the category-keyword ``supermarket". To filter retrieval results, the user can also enter the rough location of Shop A, for example, ``Shinjuku-ku, Tokyo". Because the locations of all the supermarkets visited must be indicated in the town directory database, the system accesses the town directory, and finds one or more supermarkets near his footprints including Shop A. The system then shows the user the formal names of all the supermarkets which he visited and the time of visits as retrieval results. Probably he chooses Shop A from the results. Finally, the system knows the time of the visit to Shop A, and displays the desired scene.

However, the system may make mistakes, for example, to the query shown above. Even if the user has not actually gone into Shop A but has passed in front of it, the agent will enumerate that event as one of the retrieval results.

To cope with this problem, the system investigates whether the GPS signal was received for a time following the event. If the GPS became unreceivable, it is likely that the user went into Shop A. The system investigates the length of the period when the GPS was

unreceivable, and equates that to the time spent in Shop A. If the GPS did not become unreceivable, the user most likely did not go into Shop A.

We examined the validity of this retrieval technique. First, we went to Ueno Zoological Gardens, the supermarket ``Summit", and the drug store ``Matsumoto-Kiyoshi". We found that this technique was very effective. For example, when we referred to a name-keyword ``Summit", we found the scene that was captured when the user was just about to enter ``Summit" as the result. When we referred to the category-keyword ``drug store", we found the scene that was captured when the user was just about to enter ``Matsumoto-Kiyoshi", and similarly for Ueno Zoological Gardens. These retrievals were completed quickly; and retrieval from videos for three-hours took much less than one second.

Fig. 7 Retrieval with the town directory

5. Frame Extraction based on Spatio-Temporal Sampling

The life log system captures our daily life, it contains very large amount of data and the recorded time can be very long. So, it is difficult to process and extract all visual data. In addition, content feature extraction is computationally very expensive as compared to context based information from various sensors. Features extracted form content such as color, shape and so on, are very useful and related to contexts.

Then, only the extracted key frames are subjected to sophisticated processing such as image analysis. For these reasons, we present a key frame extraction method by spatio-temporal sampling. Sampling scheme can also produce summary of the long sequence. Then, good sampling makes the user glance over the long sequence.

In spatio-temporal sampling, the location and the time of the recording are sampled using data from the GPS and time data. In addition, we make use of derivatives of the location, which define movement of the user, such as speed and direction. Changes of location and their derivatives are used for triggers for the spatio-temporal sampling. The entire volume is summarized by key frames extracted.

Extracted frames from the time data can be summarization of life log video. Key frame extraction in Figure 8 shows the frames extracted every 60 seconds. As these key frames, we can know the user are moving and imagine his situation roughly.

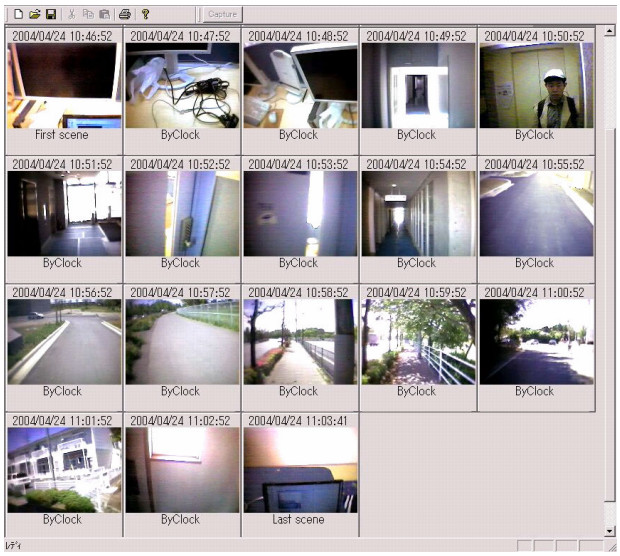

Fig. 8 Frames extracted every 60 seconds

In order to see more details, we have to extend the view of scenes by using the shorter sampling intervals. The frames sampled every 30 seconds were presented in Figure 9. Then more valid images of the traveling way can be seen as 12 key frames

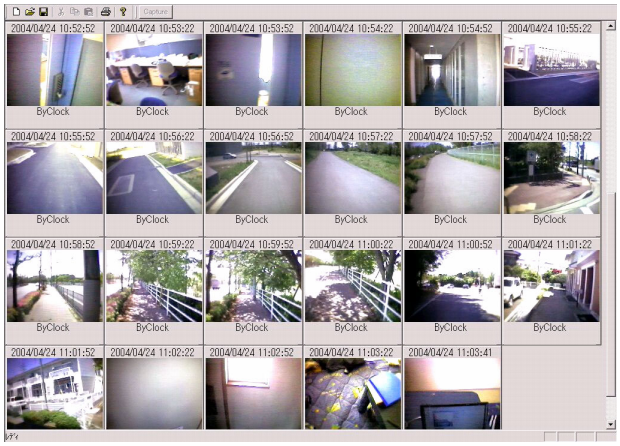

Fig. 9 Frames extracted every 30 seconds

Location is an important context that we can get directly from GPS data. We can know the approximate distance by measuring the distances between locations. Figure 10 shows the extracted frame sampled every 50 meters. In the key frames sampled by distance, we see the

summarization of views of locations. This sampling scheme is suitable for summarizing movement of the user so that we can see the different scenes in different places.

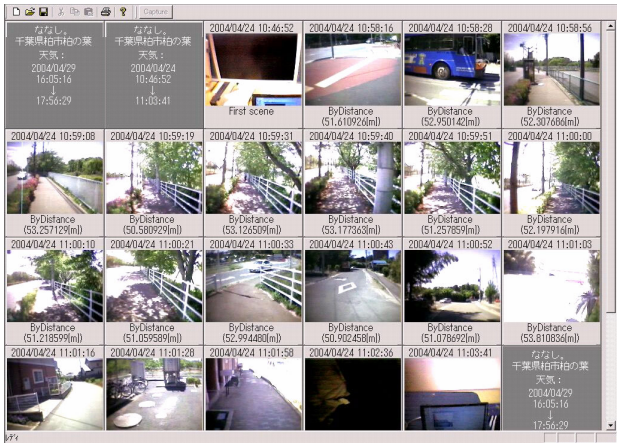

Fig. 10 Frames extracted every 50 meters

We can control sampling using speed of the movement. In case when we want to extract scenes when user stands still, we can filter the sampling scenes by speed of the movement. In Figure 11, in which the threshold of speed is set below 3 km/h, and the scenes with user's speed less than the threshold are further extracted. The scenes when user rides a bicycle are eliminated.

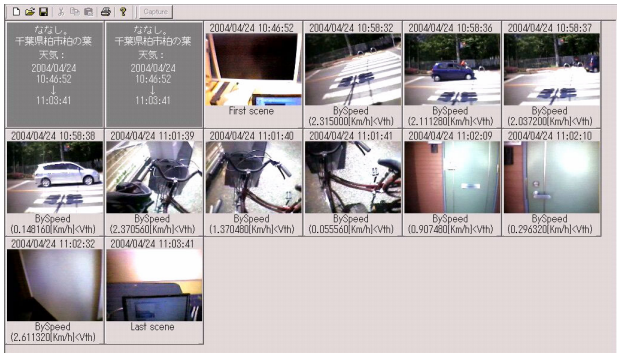

Fig. 11 Frames extracted by Speed

In the view point of navigation, the acceleration, the deceleration, and the turning are important actions. The frames shown in Figure 12 are extracted when speed or direction of movement is changed.

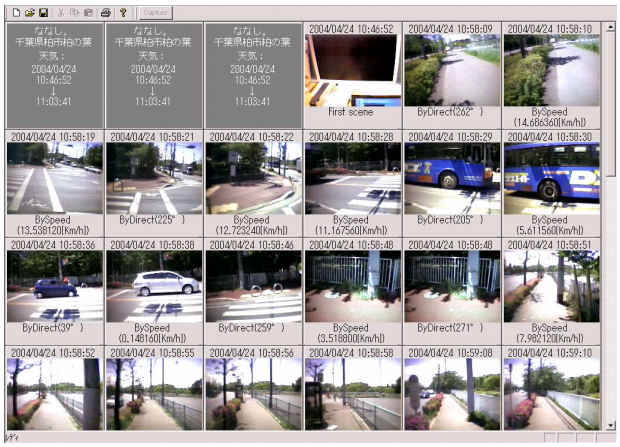

Fig. 12 Frames extracted by speed or direction.

One of the examples that key frame sampling is very useful is conversation scene detection because conversation scene contains some interesting talking topics in our life. However, audio and visual processing has to use computational load more then context one. So, it will be better if we can divide the scene by using the context.

6. Conclusions and Ongoing Works

In this paper, we presented continuous capture of our life log with various sensors and additional data, and proposed effective retrieval using context. We explained our life log capturing system which has video, audio, acceleration sensors, a gyro, GPS, annotations, documents, web pages, and emails. We also described our retrieval method based on context.

As extensions to our current system, we are working on several additional functions and directions. First, we attempt to make more use of content in addition to context. Conversation is an important key for retrieval of our daily life. We therefore attempt to detect conversation scenes. Conversation scene detection requires content analysis. We examined that we can detect to some extent conversation scenes based only on audio and video data [17].

Finally, we started to use sensor data obtained by sensors embedded in the environment. Experiments are being done in a “ubiquitous home” where a number of cameras, various sensors are installed. Integrated use of environmental sensor data together with wearable sensors is being investigated.

References

[1] M. Lamming and M. Flynn, “Forget-me-not” Intimate Computing in Support of Human Memory. Proceedings of FRIEND21, 94 Int. Symp. Next Generation Human Interface, 125-128 (1994).

[2] J. Gemmell, R. Lueder and G. Bell, Living with a Lifetime Store. ATR Workshop on Ubiquitous Experience Media, (2003).
[3] S. Mann, "WearCam" (the wearable camera): personal imaging system for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis. Proceedings of ISWC'98, 124-131 (1998).
[4] J. Healey and R.W. Picard, StartleCam: a cybernetic wearable camera. Proceedings of ISWC'98, 42-49 (1998).
[5] T.Kawamura, Y.Kono and M.Kidode, Wearable interfaces for a video diary: towards memory retrieval, exchange, and transportation, Proceedings of ISWC'02, 2002
[6] J. Farringdon and Y. Oni, Visual augmented memory (VAM), Proceedings of ISWC'00, (2000)
[7] Y. Nakamura, J. Ohde and Y. Ohta, Structuring Personal Activity Records based on Attention - Analyzing Videos from Head-mounted Camera, Proceedings of IEEE ICME2000", TD10-5, (2000)
[8] N. Kuwahara et al, Proposal of auto-event-recording on medical nursing by using wearable sensors, Proceedings of Interaction 2003 (in Japanese) (2003)
[9] R. Ueoka and M. Hirose, Experiential Recording by Wearable Computer, Proceedings of HCII 2001, pp.753-757, (2001)
[10] B. Clarkson and A. Pentland, Unsupervised Clustering of Ambulatory Audio and Video, Proceedings of ICASSP'99 (1999)
[11] K. Mase and Y. Sumi, Interaction Corpus and Experience Sharing, Proceedings of ATR Workshop on Ubiquitous Experience Media (2003)
[12] A. K. Dey et al., The conference assistant: Combining context-awareness with wearable computing, In Proceedings of ISWC 1999 (1999)
[13] K. Aizawa, K.Ishijima and M. Shiina, Summarizing wearable video. Proceedings of ICIP 2001, 398-401 (2001).
[14] Y. Sawahata, and K. Aizawa, Wearable imaging system for summarizing personal experiences. Proceedings of ICME 2003, I45-I48 (2003).
[15] T. Hori and K. Aizawa, Context-based video retrieval system for the Life Log applications. Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, 31-38 (2003).
[16] T.Hori and K. Aizawa, Capturing Life Log and Retrieval based on Context, IEEE ICME2004, (2004).
[17] D.Tancharoen, K.Aizawa, Novel concept for video retrieval in life log applications, PCM2004, (Dec. 2004)
[18] Pervasive 2004 Workshop on Memory and Sharing of Experiences 2004 (Apr.2004)
[19] ACM MM Workshop, The 1st ACM Workshop on Continuous Archival of Personal Experiences (Oct.2004)