PAPER

Ubiquitous Home: Retrieval of Experiences in a Home Environment

Gamhewage C. DE SILVA[†], **Toshihiko YAMASAKI**[†], *and* **Kiyoharu AIZAWA**[†,††], ***Members***

SUMMARY Automated capture and retrieval of experiences at home is interesting due to the wide variety and personal significance of such experiences. We present a system for retrieval and summarization of continuously captured multimedia data from Ubiquitous Home, a two-room house consisting of a large number of cameras and microphones. Data from pressure based sensors on the floor are analyzed to segment footsteps of different persons. Video and audio handover are implemented to retrieve continuous video streams corresponding to moving persons. An adaptive algorithm based on the rate of footsteps summarizes these video streams. A novel method for audio segmentation using multiple microphones is used for video retrieval based on sounds with high accuracy.

An experiment, in which a family lived in this house for twelve days, was conducted. The system was evaluated by the residents who used the system for retrieving their own experiences; we report and discuss the results.

key words: *Multimedia, Personal Experiences, Life-log, Video Retrieval, Ubiquitous Environment, Floor Sensors*

1. Introduction

Automated capture of experiences taking place at home is interesting for a number of reasons. Home is an environment where a variety of important events and experiences take place. Some of these, such as the first footsteps of a child, provide no opportunity for manual capture. Some others are so important that people have a strong desire to include themselves in the experience, rather than carry a camera and shoot photos or video. A corpus of interactions and experiences at home can provide valuable information for studies related to the design of better housing, human behavior, etc. Other prospective applications include assistance for elderly residents and aiding recollection of things that were forgotten.

Both capture and retrieval of experiences in a home-like environment is extremely difficult due to a number of reasons. Even the simplest and the smallest of the houses are partitioned into a number of rooms or regions, making it necessary to have a large number of cameras and a fair number of microphones for complete data capture. Continuous recording of data from these devices, to ensure the capture of all important experiences, results in a very large amount of data. The level of privacy differs at different places of a house, and sometimes certain regions are shared only among certain residents.

The most difficult problems, however, arise during retrieval and summarization of the captured data. The content is much less structured compared to that from any other environment. Queries for retrieval could be at very different levels of complexity, and the results can be in various levels of granularity. Some examples are shown below:

- "Show the video from the camera near the entrance to the living room, from 8:30pm to 9:00 pm, on the 1st of February, 2005"
- "What was our child doing between 5:30 and 6:30 pm. yesterday?"
- "On which date did Jeff visit us last month?"
- "How did the strawberry jam that I bought last week finish in 4 days?"

It is evident that the answers to different queries can be of different types of data. for example, the result corresponding to the first quesry is a lengthy video clip whereas a date stamp will suffice to answer the third query. Some queries require multiple modalities. For example, the second query is best answered by retrieving both video and audio making sure that the child is both seen and heard during the specified duration.

Given the large content and the state of the art of content processing algorithms, multimedia retrieval for ubiquitous environments based solely on content analysis is neither efficient nor accurate. Therefore, it is desirable to make use of supplementary data from other sensors for easier retrieval. For example, proximity sensors that get activated by human presence will remove the burden of image analysis for human detection. Since ubiquitous environments are built with infrastructure to support cameras and microphones for capture, it is relatively easy to add additional sensors to acquire such data. Domain knowledge, such as the purpose of use for each room, is also helpful in the design of algorithms for retrieval.

This article presents our work on capturing and retrieval of personal experiences in a ubiquitous environment that simulates a home. The objective is to create an electronic chronicle [1] that enables the residents of the house to retrieve the captured video using simple and interactive queries. Context data from pressure based floor sensors are analyzed using unsu-

[†]The author is with the Department of Frontier Informatics, University of Tokyo.

[††]The author is with the Department of Electronic Engineering, University of Tokyo.

pervised data mining algorithms to achieve fast and effective retrieval and summarization of video and audio data. Audio analysis and segmentation are used to complement context based retrieval. Accuracy measures and experiments are designed and conducted to evaluate the performance of the algorithms developed, and results are reported. Of particular importance are the results of a real-life experiment where a family lived in this home and used the system for retrieval of their experiences.

2. Related Work

This research combines the work from the two main research areas of Ubiquitous Environments and Multimedia Retrieval. Ubiquitous environments are equipped with a large number of sensors of different types, enabling acquisition of data regarding the events that take place in them. They are sometimes referred to as smart environments, if they are able to recognize and respond to the actions of the humans in the environments.

The current research on smart and ubiquitous environments can be divided in to two major categories. One aims at providing services to the people in the environment by detecting and recognizing their actions. Such environments serve as information appliances; examples are the Aware Home Project [2] and Ambient Intelligence Project [3]. The other aims at storing and retrieval of media, in different levels from photos to experiences. This type of research has become possible due to the recent developments in storage technologies facilitating recording large amounts of data. Some of the projects, such as CHIL [4], attempt to combine both directions by supporting user interaction real-time and using retrieval for long term support.

A discussion of the state of the art of multimedia retrieval can be found in [5]. Most of the existing research deals with a previously edited single video stream with specific content. For such data, the common approach is content analysis, making use of domain knowledge where applicable. However, the use of context data where available, can improve the performance greatly [6]. Life-log video captured by a wearable camera has been indexed and retrieved successfully by using supplementary context information such as location, motion, and time [7].

There are several ongoing projects that work on multimedia retrieval for ubiquitous environments. The Ubiquitous Sensor Room [8] is an environment that captures data from both wearable and ubiquitous sensors to retrieve video diaries related to experiences of each person in the room. Jaimes et al. [9] utilize graphical representations of important memory cues for interactive video retrieval from a ubiquitous environment. The Sensing Room [10] is a ubiquitous sensing environment equipped with cameras, floor sensors and RFID sensors for long-term analysis of daily human behavior.

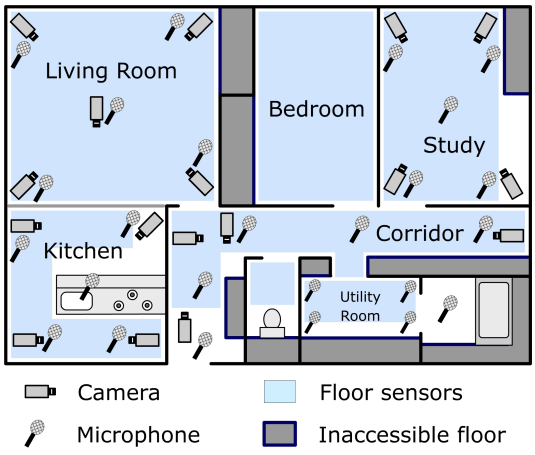

Fig. 1 Ubiquitous home sensor layout.

Video and sensor data are segmented into 10-minute intervals and the activity in the room during each segment is recognized using a Hidden Markov Model. Matsuoka et al. [11] attempt to understand and support daily activity in a house, using a single camera installed in each room and sensors attached to the floor, furniture and household appliances.

3. Ubiquitous Home

3.1 Sensors and Data Acquisition

The Ubiquitous Home [12] has been designed to provide a testing ground for ubiquitous sensing in a household environment. Figure 1 shows the sensor layout of the ubiquitous home. Simulating a two-bedroom house, it is equipped with 17 cameras and 25 microphones for continuous acquisition of video and audio. Pressure based floor sensors are mounted in the areas shown in light blue in Figure 1.

The cameras are adjustable, but stationary during capture. Images are recorded at the rate of five frames per second and stored in JPEG file format. The frame rate is low due to storage space restrictions, but given human behavior in a household environment this frame rate is adequate. Audio is sampled at 44.1 kHz from each microphone and recorded into audio clips in mp3 file format. The duration of each clip is 1 minute. Creating audio clips of short duration facilitates faster searching, by enabling the use of the file names as an index. This also saves the memory of the computer used for retrieval, by preventing loading a huge file containing a large amount of data.

The floor sensors are point-based pressure sensors spaced by 180mm in a rectangular grid. Their coordinates are specified in millimeters, starting from the bottom left corner of the house floor as seen in Figure 1. The sensors are interfaced to a hardware controller that samples the pressure on each sensor at 6 Hz. At the start of data acquisition, the sensors are initialized

to be in state '0'. When the pressure on a sensor crosses a specific threshold, it is considered to change its state to '1'. A pair of state transitions occurs when a foot is placed on and removed from a sensor. Each pair is combined to form a sensor activation, with the attributes shown in Table 1, and recorded.

A few issues arise from the construction, installation and interfacing of sensors. Given the spacing between the sensors and the average size of a human foot, a single footstep can activate between 1 to 3 sensors. Rubber damping on sensors can cause a delay in activation. This delay, combined with the low sampling rate, can occasionally miss out a footstep completely, according to manual observation of data.

3.2 Data Collection

Two types of experiments were conducted for data collection in ubiquitous home. The first type of experiments, hereafter referred to as students' experiments, were conducted by students working on research related to the ubiquitous home. Most of these experiments were aimed at acquiring training data for specific actions and events. In one of the experiments, for instance, students gathered data for different numbers of people walking along predetermined paths inside the house. In order to gather test data, two students spent three days in the ubiquitous home. Data were acquired from 9:00 a.m. to about 5:00 p.m. each day. The subjects performed simple tasks such as cooking and having meals, watching TV, and cleaning the house. They had meetings with up to five visitors at a given time, inside the ubiquitous home. The actions of the subjects were not pre-planned for this experiment. Audio data were not available during the time these experiments were conducted.

Since the experiments mentioned above do not represent life at home properly, a real-life experiment was conducted. A family of three members (a married couple with their 3-year old daughter) stayed in the ubiquitous home for 12 days. They lead their normal lives during this stay. The husband went to work on weekdays; the wife did the cooking; and everybody went out at times during the weekend. No manual monitoring of video was performed during the experiment.

The following sections describe the algorithms used for retrieval of experiences from the data collected as above. The processing and analysis were performed offline. However, the algorithms were designed in such a way that they can be adapted for real time processing.

Table 1 Format of sensor activation data.

Start Time	Duration	X	Y
2004-09-03 09:41:00.453	1.922	2100	3610
2004-09-03 09:41:20.640	1.328	1920	3250

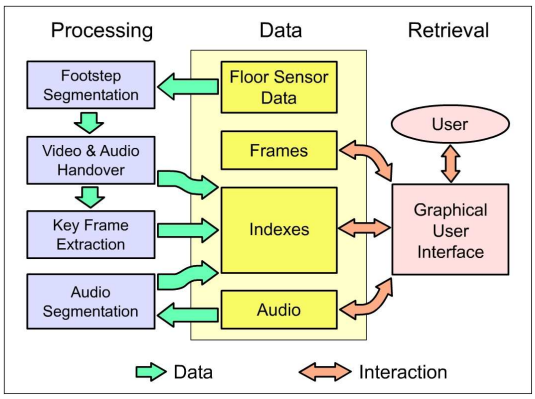

Fig. 2 Schematic of the proposed system.

4. Retrieval

The main issue in retrieval for this environment is the large amount of sources and data. Furthermore, only a few data sources will convey useful information at any given time due to the relatively small number of residents in a home and their grouped behavior. Our approach in this work is to automatically select sources that will convey the most amount of information based on context data. Only the selected sources will be queried to retrieve data and these data will be analyzed further for retrieval, thereby minimizing the need for content analysis.

Figure 2 outlines the functionality of the proposed system. We start retrieval by analyzing the floor sensor data. Unlike a video camera or a microphone that covers a limited range, floor sensors cover almost the entire house and provide data in a compact format. This makes it possible to process them faster with relatively low processing power. The results are used for extracting only the relevant portions of audio and video data to be analyzed for further retrieval. Audio is analyzed separately as a cue for video retrieval.

4.1 Footstep Segmentation

The floor sensor activation data contains two types of noise. One of these is characterized by very small durations (30-60 ms). These are likely to appear when there are footsteps on adjacent sensors. The other occurs when a relatively small weight such as a leg of a stool is placed on a sensor. The result is a series of localized sensor activations occurring periodically. We constructed Kohonen Self Organizing Maps (SOM) using the variables X, Y and duration of sensor activation data, for noise reduction. Both types of noise formed distinct clusters in SOM's, enabling easy removal.

A 3-stage Agglomerative Hierarchical Clustering (AHC) algorithm, described in our previous work [14], is used to segment sensor activations into footstep sequences of different persons. Figure 3 is a visualization

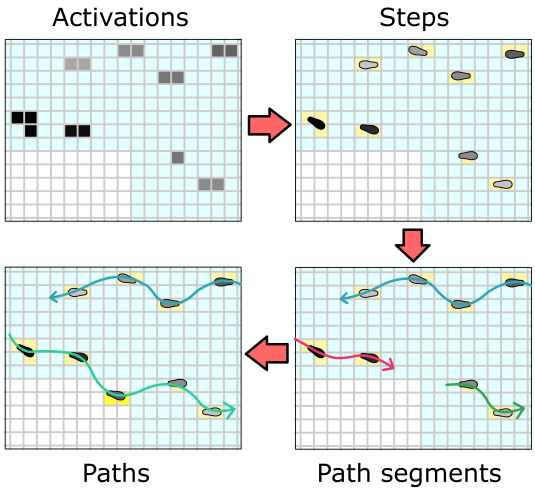

Fig. 3 Footstep segmentation.

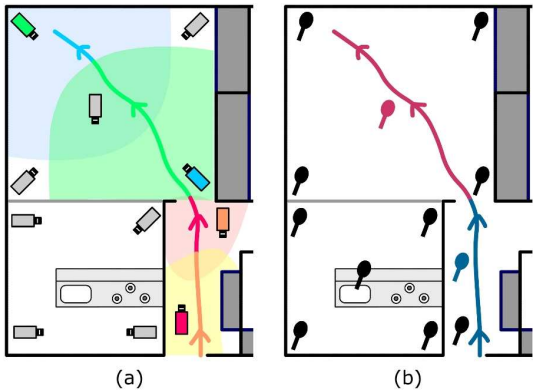

Fig. 4 An example of creating a video for a person's path. (a) Video handover, (b) Audio handover.

of this process. The grid corresponds to the resolution of floor sensors, which are shown in light blue. Activations that occurred later are indicated with a lighter shade of gray.

In the first stage, sensor activations caused by a single footstep are combined. The distance function for clustering is based on connectedness and overlap of durations. In the second stage, the footsteps are combined to form path segments using a distance function which is based on the physiological constraints of walking such as the range of distances between steps, the overlap of durations in two footsteps, and constraints on direction changes. However, due to the low resolution and the delay in sensor activations, the floor sensor data are not exactly in agreement with the actual constraints. Therefore, we obtained statistics from several data sets corresponding to a single walking person and used the statistics to identify a range of values for each constraint. The third stage compensates for the fragmentation of individual paths due to the absence of sensors in some areas, as shown in the bottom left of Fig. 3. The starting and ending timestamps of path segments, context data such as the locations of the doors and furniture and information about places where floor sensors are not installed, are used for clustering.

This algorithm performs fairly well in the presence of noise and activation delays, and despite the absence of floor sensors in some areas of the house. However, two types of errors are present in the segmented paths. Some paths are still fragmented after clustering in the third stage. There are some cases of swapping paths between two persons when they walk close to each other.

The performance of this algorithm was evaluated using a data set of approximately 27000 sensor activations, corresponding to 10 hours of data acquisition [13]. The algorithm segmented 52 paths, where the actual number of paths was 52. There were 15 instances of fragmentation and four instances of swapping paths.

4.2 Video Handover

We intend to create a video clip keeping a given person in view as he moves within the house. Since the cameras are stationary with fixed zoom, this seems trivial if footstep segmentation has been accurate. However, with more than one camera that can see a given position, it is necessary to select cameras in a way that a "good" video sequence can be constructed. The users might have their preferences, such as the minimum possible number of transitions, frontal view wherever possible, or the least amount of occlusion by others. We refer to this task as video handover.

We use position-based handover [14], an algorithm described in our previous work, for camera selection. This algorithm is based on a simple view model, where the viewable region for each camera is specified in terms of floor sensor coordinates. The main objective in this algorithm is to create a video sequence that has the minimum possible number of shots. If the person can be seen from the previous camera (if any), then that camera is selected. Otherwise, the viewable regions for the cameras are examined in a predetermined order and the first match is selected. Figure 4a demonstrates how this algorithm works. The arrow indicates the path of the person. Each shaded region on the house floor corresponds to the region viewed by the camera indicated by the same color. The change of color of the arrow indicates how the camera changes with the position of the person.

This algorithm is computationally simple, and creates video clips with camera changes that seem natural to the viewers. Despite not making any attempt to capture frontal images, it is possible to acquire a frontal view of a walking person most of the time, due to the positioning and orientation of cameras.

4.3 Audio Handover

The next step is to 'dub' the video sequences created

by video handover. Although there are a large number of microphones, it is not necessary to use all of them since a microphone can cover a larger region compared to a camera. Furthermore, frequent transitions of microphones can be annoying to listen. We implement a novel, simple algorithm for audio handover. Each camera is associated with one microphone for audio retrieval. For a camera installed in a room, audio is retrieved from the microphone that is located in the center of that room. For a camera installed in the corridor, the microphone closest to the center of the region seen by that camera is selected. This algorithm attempts to minimize transitions between microphones while maintaining a reasonable sound level. Figure 4b shows how the microphones are selected for the video clip created in the case of Fig. 4a. It was possible to create sound tracks with a reasonably uniform amplitude level, using this approach. Transitions were smooth, other than for occasional instances where a person was moving from one room to another while talking.

4.4 Key Frame Extraction

The video sequence constructed using video handover has to be sampled to extract key frames. To create a summary that is both complete and compact, we have to minimize the number of redundant key frames while ensuring that important key frames are not missed.

We designed and implemented four algorithms for key frame extraction, as summarized in Table 2. In all entries, T is a constant time interval. Temporal sampling and spatial sampling are relatively simple algorithms where key frames are sampled according to the time and the person's movement respectively. These two are combined in spatio-temporal sampling in a way that they complement each other. However, it is evident that we should try to acquire more key frames when there is more activity and vice versa. Since the rate of footsteps is an indicator of some types of activity, we hypothesize that it is possible to obtain a better set of key frames using an algorithm that is adaptive to the rate of footsteps. Adaptive spatio-temporal sampling is based on this hypothesis. When there is no camera change, the time interval for sampling the next key frame is reduced with each footstep, thereby sampling more key frames when there are more footsteps.

We designed and conducted an experiment to evaluate the performance of these algorithms [15]. Eight voluntary subjects took part in the experiment. First, the subjects extracted key frames from four video clips according to their own choice. These key frames were used to create average key frame sets, to be used as ground truth for evaluation purposes. Thereafter, they observed seven key frame sets created by the system, for each video clip, using different algorithms; a set created by spatial sampling, two sets each by the other

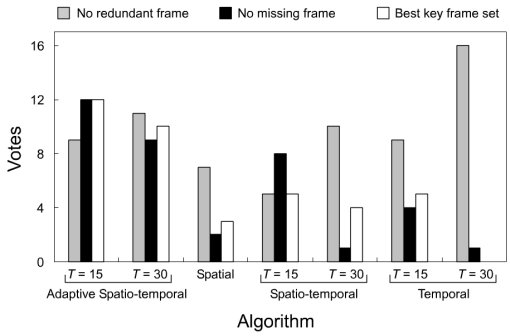

Fig. 5 Comparison of algorithms for key frame extraction.

algorithms with $T = 15$ s and $T = 30$ s. The subjects ranked these frame sets in terms of the number of key frames, the number of redundant frames and the number of missing frames. They voted for the key frame set that summarized the sequence best, justified their selection and suggested improvements.

Figure 5 summarizes the results of this evaluation. Adaptive spatio-temporal sampling performed much better than the other algorithms and 15 s was found to be the better value out of the two, for the parameter T .

The key frame sets extracted by the system were evaluated both subjectively and quantitatively, using the average key frame sets. Figures 6a and 6b show the average key frames and the frame set created by adaptive spatio-temporal sampling respectively, for one sequence. Figure 6c shows the path of the person in the sequence, with her locations when the key frames were sampled. The algorithm failed to capture the key frame corresponding to the girl picking a camera from the stool (bottom right of Fig. 6a). It extracted two

Table 2 Algorithms for key frame extraction

Sampling algorithm	Condition for sampling key frame
Spatial	At every camera change
Temporal	Once every T seconds
spatio-temporal	Sample a key frame • at every camera change • If T seconds elapsed with no camera change after the previous key frame
Adaptive spatio-temporal	Sample a key frame • at every camera change • if t seconds elapsed without a camera change where: $t = T(1 - n/20)$ if $1 \leq n \leq 10$ $t = T/2$ if $n > 10$ (n = number of footsteps since last key frame)

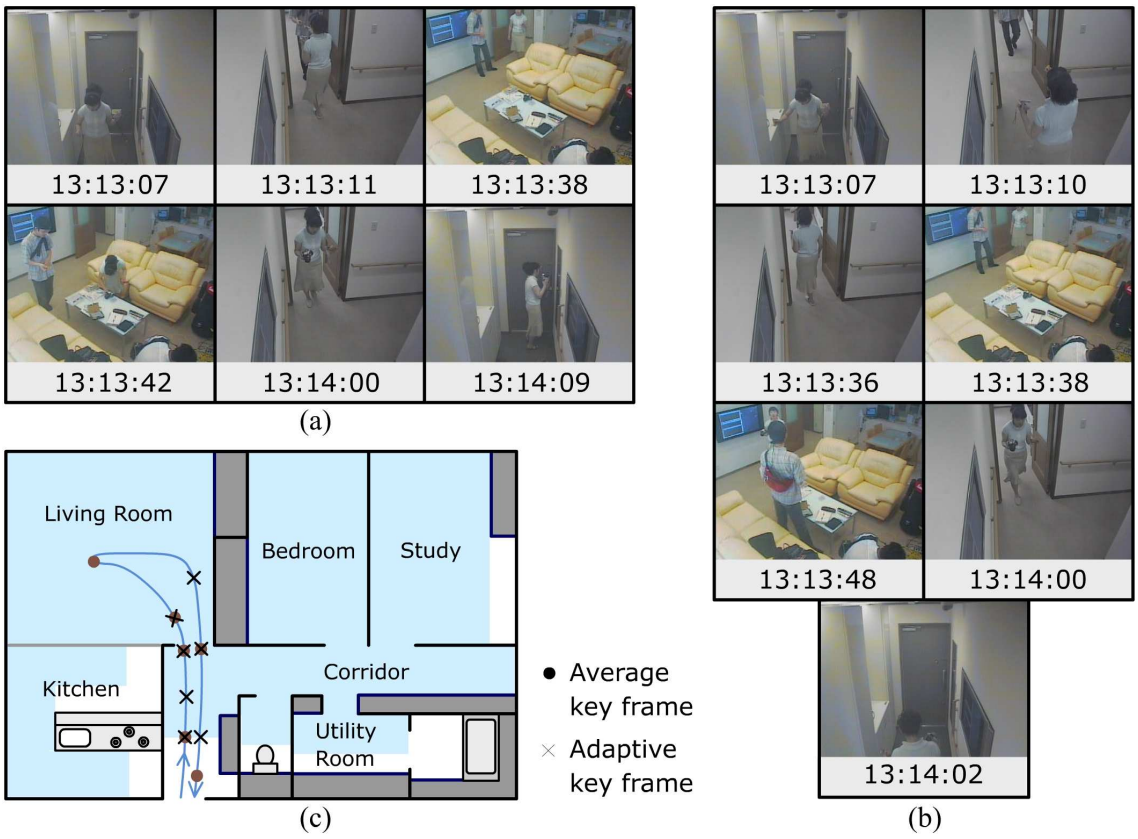

Fig. 6 Subjective evaluation of key frame extraction. (a) average key frames, (b) key frames extracted by adaptive spatio-temporal sampling, (c) path of the person with the positioning of key frames.

redundant frames as she was within the same view for a longer time. Quantitative evaluation based on average sets of key frames showed that approximately 80% of the most desired key frames can be retrieved using adaptive spatio-temporal sampling.

4.5 Audio Segmentation for Retrieval

The floor sensors are unable to capture data when the people are not treading on a floor area with sensors. Furthermore, they are not activated if the pressure on the sensors is not sufficiently large: for example, when a person is sitting and leaning back with the feet resting on the floor. Audio data can be used to supplement video retrieval in such situations. Audio-based retrieval can also be conducted independently, to support various types of queries.

With 25 microphones recording continuously, the amount of audio to be processed is quite large. Redundancy in data is high since the microphones are located in close proximity. A trade-off has to be made between utilizing the redundancy to improve the accuracy of retrieval and minimizing processing by removing redundancy.

We intend to eliminate audio corresponding to silence in each room of the house. The result is a set of audio segments that can be analyzed further or directly used to retrieve video for situations where there were some sounds at a given location. A common approach for silence elimination for a single audio stream is to compare the RMS power of the audio signal against a threshold value [16]. We select this method as the basis of audio segmentation in this work. The threshold for each microphone is estimated by analyzing audio data for silence and noise for that microphone. Audio clips with a total duration of one hour were extracted from different times of day. These clips were partitioned into frames having 300 samples. Adjacent frames had a 50% overlap. The RMS value of each frame is calculated and recorded, and the statistics obtained for each clip.

Since the probability distribution of the RMS values for different audio clips were not significantly different, the data were combined to make a single probabilistic model for silence and noise. The threshold value was selected to be at 99% level of confidence according to this distribution. The value was selected below 100% as false negatives (sound misclassified as silence) are more costly than false positives (silence misclassified as sound). The latter can be eliminated using further analysis.

The first stage of silence elimination is based on individual microphones. The audio stream is divided into overlapping frames in the same manner as described previously, and the RMS value of each frame is calculated. If the calculated RMS value is larger than the threshold, the frame is considered to contain sound. Otherwise, it is considered to contain silence. The total duration of contiguous frames with RMS above the threshold are used for further processing. Sets of contiguous frames with a duration less than 0.1 s are removed. Sets of contiguous frames that are less than 0.5s apart are combined together to form single segments. This method was evaluated using 18 hours of video on each microphone, and it was possible to achieve silence elimination with 0% false negatives and approximately 2% false positives. The amount of voice activity captured by each microphone varied from 14 to 400 minutes. There was no significant difference of accuracy over different microphones.

The second stage uses the data from multiple microphones in close proximity to reduce false positives. For each microphone, a binary sound segment function $B(n)$ and cumulative sound segment function $C(n)$ are defined by

$B(n) = 1$ if there is sound in the n^{th} second of audio stream

$B(n) = 0$ otherwise

for the set of microphones in the same room.

Noise is random, and usually has a small duration. Due to its randomness, it is less likely that noise in sound segments from different microphones occur simultaneously. Due to the small duration, they can be distinguished in most situations. Based on the above arguments, we use the following voting algorithm to determine the sound segment function, $S(n)$.

$S(n) = 1$ if $C(n) \otimes M(n) \geq \lceil k/2 \rceil$

$S(n) = 0$ otherwise

where $\otimes$ denotes convolution,

$M(n) = [111]$

and

$k =$ no. of microphones installed in the location

The algorithm was tested using 90 hours of audio data. It was possible to remove 83% of the false positives that remained after silence elimination in individual audio streams. The remaining noise amounted to 583 segments (approximately 10 minutes).

At the current state, the method of audio-based retrieval is fairly simple. Video is retrieved from all cameras in the room for each sound segment. In addition to this, video retrieved for footstep sequences are extended using sound segments, as follows. If there is only one footstep sequence overlapping partially with a sound segment, it is combined with the footstep sequence.

The video created by handover is extended to include the time during which sounds were present before the start of the footstep sequence, or after the sequence ended. This improves the video in certain situations, such as when a person enters a region without floor sensors but continues to talk.

5. User Interaction

The results can be retrieved by submitting interactive queries through a graphical user interface. A query is initiated by entering the time interval for which the summary is required. The next step is to specify whether retrieval should be based on persons (video clips and key frames showing tracking a person as he moves) or locations (video clips retrieved from a location specified by user).

The initial result of person-based retrieval is a summary consisting of key frames with timestamps. For people who entered or left the house during the time interval, the key frames showing those entering or leaving the house will be displayed with timestamps. For those who entered the house before the specified time interval and remained inside, a key frame at the start of the time interval is displayed. By clicking each key frame, it is possible to retrieve a video clip or a set of key frames showing the person appearing in the key frame. Figure 7 demonstrates interaction with the system for retrieval of key frames to summarize the behavior of a single person.

For retrieval by location, the user can click on the floor plan of the house and select cameras interactively. It is possible to watch the complete video from the selected camera for the specified duration, or filter the results so that video clips are retrieved only for occasions when sound and/or floor sensor data are present. A demonstration of how the system can be used for the retrieval of video and key frames can be found in [17].

6. User Study: Real-life Experiment

We designed and conducted a user study, where the residents for the real-life experiment studied the system, with the following objectives:

- Identify requirements for experience retrieval in the ubiquitous home, using feedback from people who lived in the environment
- Evaluate algorithms that have been implemented so far for video summarization and retrieval
- Identify directions for future work and improvement of the existing algorithms

6.1 Procedure

This study consisted of two parts; a requirements analysis for home experience retrieval, and a hands-on session of the system we developed. Data captured during six hours on the 12^{th} of April 2005, amounting to a total of 102 hours of video and 150 hours of audio data,

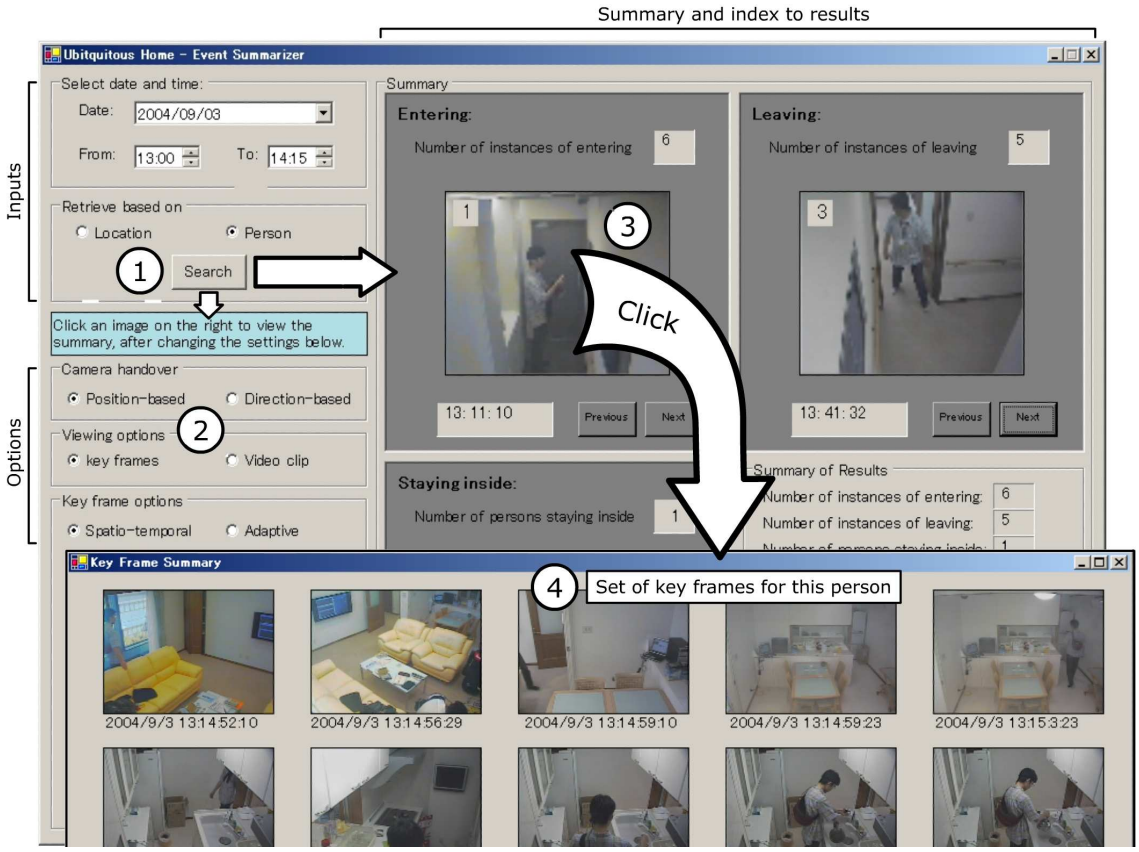

Fig. 7 User interaction with the system.

were used for the study.

The experiment consisted of three sections. The first section was a requirement study, where the subjects answered a questionnaire to specify what they would expect from a system for retrieving experiences at home. This section of the experiment was conducted before demonstrating the system, to ensure that the user requirements are not influenced by the functionality of the existing system.

In the second section, the subjects were given a demonstration on how to use the system. Only one example for each type of retrieval was shown. Thereafter, they were allowed to use the system themselves, submitting their own queries to retrieve their experiences. The authors were available in case the subjects needed advice, but were not involved in using the system. The subjects were asked to select video clips that they would like to keep. This was done both as a factor of motivation and also to find out what kind of experiences generate interest in keeping a permanent record. After using the system, the subjects rated the usability of the system by answering a brief questionnaire based on the guidelines by Chin et al.[18].

In the third section, the subjects provided descriptive feedback about the system. The subjects were asked to suggest additional requirements to what they proposed previously, in case if there were any.

The user study took approximately 3 hours, and the subjects were paid for participation. The subjects provided their responses in separate answer sheets but used the system together. This helped to elicit more responses, rather than getting only those both subjects agreed upon. Since the child is only 3 years old, only the parents actively participated in the experiment. Other than for restarting the system due to an operating system crash, no assistance was needed from the authors.

6.2 Results

The questions and the responses for the requirement study are stated below. The number of subjects who provided each answer is stated in parentheses.

1. Suppose it is possible to retrieve any event that happened anytime during your stay in ubiquitous home. What are the things you would like to see from that stay?
   - Things that I did (2)
   - Things that the other family members did (2)
   - How my child was playing when she was alone (2)
   - Things that I have forgotten (2)
   - Things we did together (1)

- A summary of what I did each day (1)

The subjects added the following after using the system.

- Recall what we did when my friend visited the house (1)
- See my own behaviour and habits, e.g.: gait (1)
- See my child growing up over a long period of time (1)

2. Supposing it is possible to have the same facility at your home, and only your family has access to the data:
   a. What would you like to use it for?
      - For taking care of my old mother; check whether she took medicine properly, or she ate too much sweets, etc. (1)
      - For finding lost objects, discover our own habits, find out how the child is behaving so that anything bad can be corrected (1)
   b. How would you like to record the data?
      - Record everything from daily life (2)
   c. Which parts of the house will you record everything?
      - Non-private areas of the house, like the living room, the kitchen, etc. (2)
      - Child's room (1)
   d. Which parts of the house and times will you refrain from recording?
      - Private places, such as bedrooms (2)
   e. Do you have any other preferences, such as times of day, about recording?
      - I would like to see my child during day time and afternoon, when I am not at home(1)
      - I want to record leisurely times playing with the child (1)
      - I Would like to record busy hours of the day for discovering things that were left behind (1)

For the usability assessment, the following were the responses from the two subjects on a seven-point response scale with 1 being the worst rating and 7 being the best.

- Learning to use the system - 6,6
- Ease of using the system - 4,5
- Overall impression - 5,6

The following are the questions in the section for feedback about the system, and the responses from the subjects.

1. How much did you remember from what you could see in the video and key frames?
   - There were many things that I did not remember. For example, that I worked that day (1)
   - I did not remember much of what happened on that day. But I was able to get a continuous recollection of the day after watching the video (1)
2. Was it possible to see interesting things that you did not see/know before?
   - Yes (2)
   - We could discover things like how our child woke up in the morning (1)
   - I was surprised to learn that I spend so much time with my child (1)
3. Out of what you saw, which parts of the video would you like to keep with you?
   - Video of the child (2)
   - Video of activities we did together, such as having meals (1)
4. State what you like about this software.
   - Automatic camera change (1)
   - Ability to see what happened when I was away (1)
5. State what you don't like about it.
   - It might reveal things that are not nice to know (1)
   - Too many video clips and key frames to look at (1)
6. For what kind of things will this software be useful to you?
   - Family diary, security
   - Taking care of family members
   - Record of our child's life
   - See myself objectively.

The responses to the requirements show that the system can already match most of the requirements the subjects had in their mind before using it. The subjects found the system easy to use, as suggested by the high rankings for the usability assessment. Descriptive feedback indicates that the subjects found the software useful and it helped them to discover a few things that they were not conscious about or did not know at all.

The subjects managed to recall what happened in the entire session and to retrieve video they wanted to watch, by using the system. They found two types of video more interesting, and watched them repeatedly. One type contained video of the child when she was alone: an example was the video clip created when the child woke up in the morning, found that she was alone in the living room, and ran for the mother. The

other type corresponded to activities that they did together, such as taking meals and playing with the child. They requested copies of both these types of video clips. The subjects used key frames as an index to the original video, rather than viewing only the key frames as a summary. They liked using the system, and it was somewhat difficult to get them to stop watching videos and answer the questionnaire.

7. Discussion

7.1 Issues Related to Capture

The current system is based on continuous capture of sensor data and retrieval of significant events. An alternative is to perform *active capture*, where recording of audio and video data will start upon the detection of actions and events. This will also result in a reduction of the disk space requirement for capture. However, there are two main reasons behind the choice of continuous capture. One, the research was carried out at a different location from the home-like environment. Therefore, it is necessary to perform a continuous capture first, and process the data offline to verify that the actual events are retrieved with sufficient precision and recall. The other reason is that experiments with families are quite difficult to arrange and the cost of losing important data due to algorithms with insufficient accuracy is quite high. However, the algorithms can be adopted to facilitate active capture in a future version of the system.

The main problem in continuous capture is the large amount of disk space consumed. Data acquisition itself requires some CPU power, for digitizing data. The video data are stored as frames and the audio as clips of 1 minute, for faster access. This results in low compression of video and fragmentation of disk space. For example, the size on disk for data captured during a single day is about 500GB, while the actual total file size is only about 200 GB. Although fragmentation is not a big issue as it can be removed, improved techniques for compression and storage will be necessary for continuous data acquisition for a long time.

The number of cameras and their positioning ensure every location of the house, unless excluded deliberately, is captured. However, some of the microphones seem to be redundant, given their range and directivity. Although we were able to use redundancy effectively in audio segmentation, it may still be possible to record from the minimum possible number of microphones to save disk space.

The floor sensors facilitate tracking people with less computational effort compared to using image analysis, where calibration and occlusion handling is necessary to achieve similar precision. While they perform more accurately than ultrasonic tracking systems and infra-red motion sensors, they are relatively more expensive and difficult to maintain. However, movement of furniture causes superfluous data, making tracking difficult. According to the results of footstep segmentation, furniture contributed to about 30% of the errors.

7.2 Algorithms for Retrieval

The accuracy of footstep segmentation deteriorates when the number of persons in the house is large and with the movement of furniture. Furthermore, accurate segmentation is not feasible when two or more persons enter a region without floor sensors. Although it is possible to improve accuracy by considering future footsteps when segmenting a given footstep, this renders the algorithm unusable in real-time.

Video handover can be improved by considering occlusion by other persons when selecting the camera. For audio handover, smoother transitions are possible by looking for silence near the point of microphone change.

The subjects who took part in the evaluation experiments for key frame extraction suggested that human-human and human-object interaction should be included in the extracted key frame sets [5]. While this is not completely feasible using only the floor sensor data, this should be attempted using content analysis.

The approach for audio-based video retrieval will retrieve false results if the house is located at a place where loud sounds can enter the house from outside. And example is sounds from trains if the house is located near a railway track. In such cases, further processing based on frequency or Cepstrum domain features will be necessary to identify and eliminate such sounds.

7.3 Real-life Experiment

The behaviors of residents in the two types of experiments were significantly different. While the subjects in students' experiments were independent in their actions, the behavior of the family in the real-life experiment was in the form of a group. This affected the quality of the results, too. For student experiments, video clips and summaries resulting from handover and key frame extraction were mostly exclusive whereas those created during the real life experiment had a lot of overlap and redundancy, due to behavior as a group. For instance, when the child was following or walking by the side of a parent, the personalized video created for the child and the parent have near 100% overlap, which results in redundancy. Therefore, video retrieval for group behavior seems to be more important for a real-life situation. Furthermore, the accuracy of footstep segmentation decreases because of complex walking patterns created by a child walking with a parent. The accuracy is about 30% less than reported previously for students' experiments.

With only two persons actively taking part in the user study, the responses have little statistical value. However, their keen interest on using the system and positive feedback justifies the motivation and the current progress of this work. The responses also provide valuable insights to identify further requirements and possible improvements. Continuing further study with other families, as the system is being developed, will help the system to evolve into one that is very useful. Experiments with longer durations will be necessary to assess the psychological effects and mental burden on residents about their life being recorded, and to investigate the commercial feasibility of such homes.

As the subjects indicated in their feedback, privacy of the residents should be protected by recording data only in the public locations of the house. Although this reduces the ability of the system to function as a memory assistant, it is an important measure as individual privacy is important even for the members of the same family. Furthermore, the system was helpful for the residents even with restrictions in locations. It can be suggested that one of the reasons for the success of the real-life experiment (in the sense that the residents enjoyed their stay and retrieval of their experiences) is that the residents were not confined to the house, and their privacy was protected even when they were in the house.

8. Conclusion and Future Work

We have implemented video retrieval and summarization for a home with a large number of sensors, by analyzing signals from pressure based sensors mounted on the floor. Hierarchical clustering followed by video handover enabled the creation of personalized video clips using a large number of cameras. It was possible to dub this video with reasonably good quality, using audio handover. An adaptive algorithm enabled retrieval of more than 80% of the key frames required for a complete summary of the video. Basic audio analysis enabled accurate audio segmentation and thereby enhanced retrieval. The residents who evaluated the system found it useful, enjoyed using it and provided valuable feedback in improving the system.

Future work will focus on further clustering of floor sensor data and classification of audio data. These will lead to the detection and recognition of higher-level actions and events, such as conversations, thereby enhancing the functionality of the system. Face detection in retrieved images and video can provide additional information for searching within the data. Novel techniques for user interaction and visualization of results will be designed, to achieve more effective and efficient retrieval. We consider comments and feedback from the residents as very important, as usability is one of the most important criteria for an effective retrieval system.

Acknowledgment

The authors would like to thank Dr. Hirotada Ueda and Dr. Tatsuya Yamasaki for their support in acquiring data from ubiquitous home, and Mr.Atsushi Omiya and family for partiticating in the real-life experiment.

References

[1] Pingali, G., Jain, R. Electronic Chronicles: Empowering Individuals, Groups, and Organizations Proceedings of IEEE International Conference on Multimedia and Expo (ICME) 2005.

[2] Abowd, G. D., Bobick, I., Essa, I., Mynatt, E., Rogers, W. The Aware Home: Developing Technologies for Successful Aging. In proceedings of American Assoc. of Artificial Intelligence (AAAI) Conf, 2002.

[3] Philips Research. Ambient Intelligence: changing lives for the better. http://www.research.philips.com/technologies/syst-softw/ami/background.html, Koninklijke Philips Electronics N.V., 2005.

[4] Waibel, A. CHIL - Computers in the Human Interaction Loop. In proceedings of LEARNTEC 2005.

[5] Jaimes, A., Christel, M., Gilles, S., Sarukkai, R., Ma, W. Y. Multimedia Information Retrieval: What is it, and why isn't anyone using it? ACM MIR 2005.

[6] Davis, M. King, S. Good, N. From Context to Content: Leveraging Context to Infer Media Metadata. ACM Multimedia 2004, 188-195.

[7] Aizawa, K., Kawasaki, S., Tancharoen, D., Yamasaki, T. Efficient Retrieval of Life Log based on Context and Content. ACM CARPE, 2004.

[8] Department of Sensory Media - Ubiquitous Sensor Room, http://www.mis.atr.jp/ megumu/IM_ Web/MisIM-E.html, ATR Media Information Science Laboratories, Kyoto, Japan.

[9] Jaimes, A. Omura, K., Nagamine, T., Hirata, K. Memory Cues for Meeting Video Retrieval. CARPE 2004, 74-85.

[10] Mori, T., Noguchi, H., Takada, A., Sato, T. Sensing Room: Distributed Sensor Environment for Measurement of Human Daily Behavior. First International Workshop on Networked Sensing Systems (INSS2004), 40-43.

[11] Matsuoka, K., Fukushima, K. Understanding of Living Activity in a House for Real-time Life Support. SCIS & ISIS 2004, 1-6.

[12] Yamazaki, T. Ubiquitous Home: Real-life Testbed for Home Context-Aware Service. Tridentcom2005, 54-59.

[13] De Silva, G. C., Ishikawa, T., Yamasaki, T., Aizawa, K. Person Tracking and Multi-camera Video Retrieval Using Floor Sensors in a Ubiquitous Environment. CIVR 2005.

[14] De Silva, G. C., yamasaki, T. Ishikawa, T., Aizawa, K. Video Handover for Retrieval in a Ubiquitous Environment Using Floor Sensor Data. ICME 2005.

[15] De Silva, G. C., Yamasaki, T., Aizawa, K. Evaluation of Video Summarization for a Large Number of Cameras in Ubiquitous Home. Proc. ACM Multimedia 2005. p. 820-828.

[16] Liu, M., Wan, C. A Study of Content-Based Classification and Retrieval of Audio Database. Proc. IDEAS 2001. p.339-345.

[17] ACM Multimedia 2005 Video Program. Video Figure: Video Summarization for a Ubiquitous Home, http://www.sigmm.org/apache/video2004/resources/videos/2005/VF_ 1.mpg, ACM SigMM, 2005.

[18] Chin, J. P., Diehl V. A., Norman, K. L. Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. Proceedings of ACM CHI'88 Conference on Human Factors in Computing Systems, 1988 p.213-218.