Evaluation of Video Summarization for a Large Number of Cameras in *Ubiquitous Home*

Gamhewage C. de Silva, Toshihiko Yamasaki, Kiyoharu Aizawa
Department of Frontier Informatics, the University of Tokyo
(chamds, yamasaki, aizawa)@hal.k.u-tokyo.ac.jp

ABSTRACT
A system for video summarization in a ubiquitous environment is presented. Data from pressure-based floor sensors are clustered to segment footsteps of different persons. Video handover has been implemented to retrieve a continuous video showing a person moving in the environment. Several methods for extracting key frames from the resulting video sequences have been implemented, and evaluated by experiments. It was found that most of the key frames the human subjects desire to see could be retrieved using an adaptive algorithm based on camera changes and the number of footsteps within the view of the same camera. The system consists of a graphical user interface that can be used to retrieve video summaries interactively using simple queries.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *video*.

General Terms
Algorithms, Experimentation, Human Factors

Keywords
Video Summarization, Key Frames, Ubiquitous Environment, Floor Sensors

1. INTRODUCTION
There has been a growing interest in automated systems for video summarization, indexing and retrieval during the past few years. A fundamental step in these systems is to extract *key frames* to represent the major content of the video sequence. Extracted key frames can provide a compact representation of the video sequence, and can be used for indexing and browsing the large volume of video in an efficient manner.

MM'05, November 6–12, 2005, Singapore.

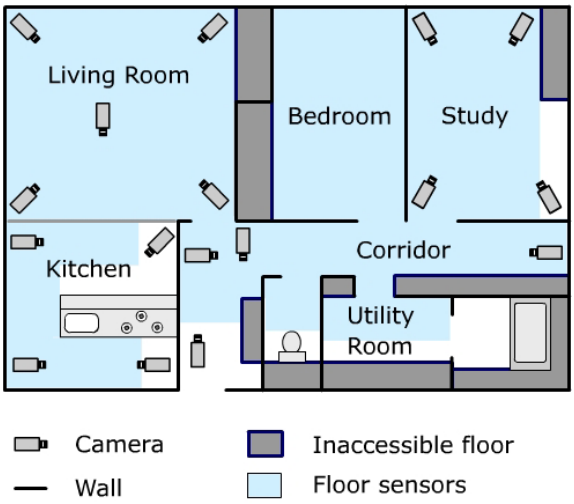

Figure 1. Ubiquitous home layout.

Video summarization for ubiquitous environments is an important task with several applications such as surveillance, creating personalized video, and aiding recollection of things that were forgotten. This is more challenging than summarization of single-stream broadcast video, which is the input for most of the existing video summarization systems. The content, usually multi-stream, is larger and less structured. When multiple cameras can see a particular location, view selection for a particular event becomes an additional issue.

Given the large amount of image data and the current state of the art of image processing algorithms, video summarization for ubiquitous environments based solely on image analysis is neither efficient nor accurate. Therefore, it is desirable to make use of supplementary data from other sensors for easier retrieval.

This research is based on *Ubiquitous Home* [1], a two-bedroom house equipped with 17 stationary cameras. This has been built to provide a testing ground for ubiquitous sensing in a household environment. Pressure-based sensors mounted on the floor, spaced by 180 mm on a square grid, are activated as people move inside the house. Video data from the cameras are continuously acquired at 5 frames per second. Figure 1 shows the sensor layout of the ubiquitous home.

By adjusting the pan, tilt and zoom of the large number of cameras, it is possible to capture every location of the house in video, ensuring that the behavior of a person in the house is fully

recorded. Monitoring a location is possible by looking at the video from the appropriate camera/s. However, personalized video retrieval and summarization for this environment is extremely tedious, if performed manually. For example, suppose Mrs. Sato wants to see what her son, Takeshi, did on the day he visited Ubiquitous Home. She remembers that Takeshi entered the house some time after 10:00 am and left the house before 12:00 noon to have lunch. In this case, it is necessary to watch the video from the camera showing the entrance to the house from 10:00 am, until the frames showing him entering the house are detected. Thereafter, it is necessary to switch between several cameras to track him as he moves within the house. The task becomes extremely tedious if the time interval for search is larger.

Our objective is to automate personalized video retrieval and summarization, for the Ubiquitous Home. We would like to create a system where the summary for the above scenario can be retrieved as follows: first Mrs. Sato enters the date and the time interval (10:00 am to 12:00 noon). The result is a set of key frames showing people who had been inside the house during this time interval. For the people who entered or left the house during the given time interval, the key frames showing them 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 browsing only the key frames showing the persons entering the house, Mrs. Sato can find the key frame showing Takeshi. By clicking on the frame, she can see a video clip or a set of key frames, showing what Takeshi did inside the house. The cameras will be selected automatically as Takeshi moves, to ensure that he appears in the video or key frames throughout his stay in the house.

Our approach to developing the above system is as follows. We analyze the floor sensor data to extract video sequences and key frames for each person in the house. The floor sensor data are much smaller compared to the large quantity of image data, making it possible to process them in real-time with relatively low processing power. We also provide an interactive interface to browse through the video and key frame summaries. To ensure that the method we use extracts key frames without missing any important actions or events while minimizing redundancy, we implement a number of methods and conduct an evaluation experiment. By defining accuracy measures and applying them on the results obtained using each method, we intend to find out which of the methods are more suitable and how they can be improved to achieve better results.

An outline of the paper is as follows: section 2 contains a review of related work; Section 3 describes the algorithms used in this paper to (a) segment footsteps of different persons in the ubiquitous home (b) create video sequences showing each person's movement (c) extract key frames from the created sequences (d) present the results to the user; Section 4 describes the setup and procedure of our experiment for evaluation of key frame extraction; Section 5 presents the results of this experiment; a discussion of these results and user comments is contained in section 6; finally section 7 concludes the paper with some suggestions for improvement and further study.

2. RELATED WORK

A thorough review of the state of the art of image and video retrieval can be found in [2]. Most related work deals with a previously edited single video stream with specific content [3] [4]. Audio is the most common supplementary input for retrieval [5] [6]. However, the use of context data where available can improve the performance greatly [7]. Life log video captured by a wearable camera has been indexed and retrieved by using supplementary context information such as location, motion, and time [8].

The *Ubiquitous Sensor Room* [9] 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. In *Aware Home* [10], floor sensors are mounted in strategic locations of the house for person identification using step signatures [11]. Jaimes et al. [12] utilize graphical representations of important memory cues for interactive video retrieval from a ubiquitous environment. The *Sensing Room* [13] is a ubiquitous sensing environment equipped with cameras, floor sensors and RFID sensors for long-term analysis of daily human behavior. 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. [14] 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.

A brief review of algorithms for key frame extraction can be found in [15]. The most common approach is *shot-based video segmentation*. First a video sequence is partitioned into a set of *shots*. A shot is an unbroken sequence of frames from one perspective. From each shot, a single key frame that represents the shot best is extracted by analyzing the image features within the shot. While this approach results in a compact representation of broadcast quality video edited by professionals, it is less effective on unedited video and video recorded by inexperienced cameramen. Moreover, this approach is not applicable to wearable video since the perspective keeps changing with the movement of the person wearing the camera.

Most of the earlier work in key frame extraction was based solely on the content. Audio has been used effectively as a supplementary input together with frame features. Aizawa et al., in their *Life-log* system, extract key frames from wearable video, based on supplementary information from several sensors [8].

Evaluation of video retrieval is a relatively new research topic. The TRECVID benchmarks, created in 2001 by the National Institute for Standardization Technology, USA, has evolved since then to include complex tasks such as *concept detection* [16]. However, TRECVID has not evaluated key frame extraction at the time of writing this paper. TRECVID 2005 [17] provides key frames at shot boundaries, but the aim is evaluation of shot detection. Kawasaki et al. [18] evaluate key frame extraction based on the percentage of unique frames. Some other researches evaluate key frame extraction indirectly where the key frames are further processed to obtain higher-level results. Song et al. [19] uses the accuracy of object segmentation from key frames as an accuracy measure for key frame extraction. However, a direct evaluation of key frame extraction would enable a more accurate comparison of the algorithms.

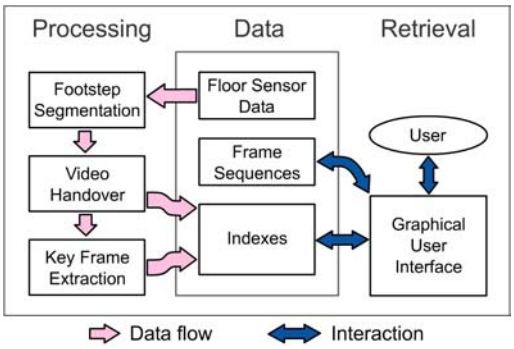

Figure 2. Schematic of the video summarizing system.

3. SYSTEM DESCRIPTION

Figure 2 outlines the functionality of the proposed system. After preprocessing the floor sensor data, footstep segmentation is carried out. This is followed by *video handover* to create a video sequence for each person. Thereafter, key frames are extracted from the resulting sequences. Indexes are created and stored in a database so that the results can be queried interactively through the graphical user interface. The following paragraphs describe each stage in detail.

3.1 Footstep Segmentation

The pressure on each floor sensor is sampled at 6 Hz. The sensors are initialized to be in state '0'. When the pressure on a sensor crosses a specific threshold, its state changes to '1'. The state transitions are recorded with the timestamp, sensor coordinates and the new state as attributes. The placement and removal of a foot on the floor results in a pair of state transitions on one to four sensors. These pairs are combined to produce data entries referred to as *sensor activations*, with attributes shown in Table 1. These sensor activations are the input for footstep segmentation.

A 3-stage Agglomerative Hierarchical Clustering (AHC) algorithm, described in our previous work [20], is used to segment sensor activations into footstep sequences of different persons. Figure 3 is a visualization of this process. The grid corresponds to the floor sensors. 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 is based on connectedness and overlap of durations. For the second stage, the distance function 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 a data set corresponding to a single walking person and used the statistics to

Table 1. Format of sensor activation data

Start Time	End Time	X	Y
2004-09-03 09:41:20.14	2004-09-03 09:41:20.96	1920	3250
2004-09-03 09:41:20.96	2004-09-03 09:41:21.60	2100	3250

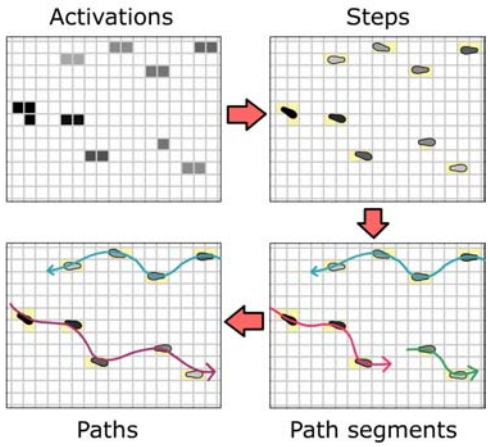

Figure 3. Footstep segmentation.

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, long steps etc. 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 well in the presence of noise and activation delays, and despite the absence of floor sensors in some areas of the house.

3.2 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. We refer to this task as *video handover*. We used *position-based handover* [21], an algorithm developed in our previous work. The algorithm is outlined in the following paragraphs.

A simple view model was constructed for each camera. The visibility of a human standing at the location of each floor sensor, through this camera, is represented by the value of 1. This mapping was created manually by observing images obtained during experiments. The set of models can be looked up to identify cameras that can see a person at a given position.

In position-based handover, the main objective 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 mappings for the cameras are examined in a predetermined order and the first match is selected.

3.3 Key Frame Extraction

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

A simple approach to summarize the video is *temporal sampling*, i.e. sampling key frames periodically. However, this algorithm extracts a large number of redundant key frames if there is little activity or the sampling interval is too small. On the other hand,

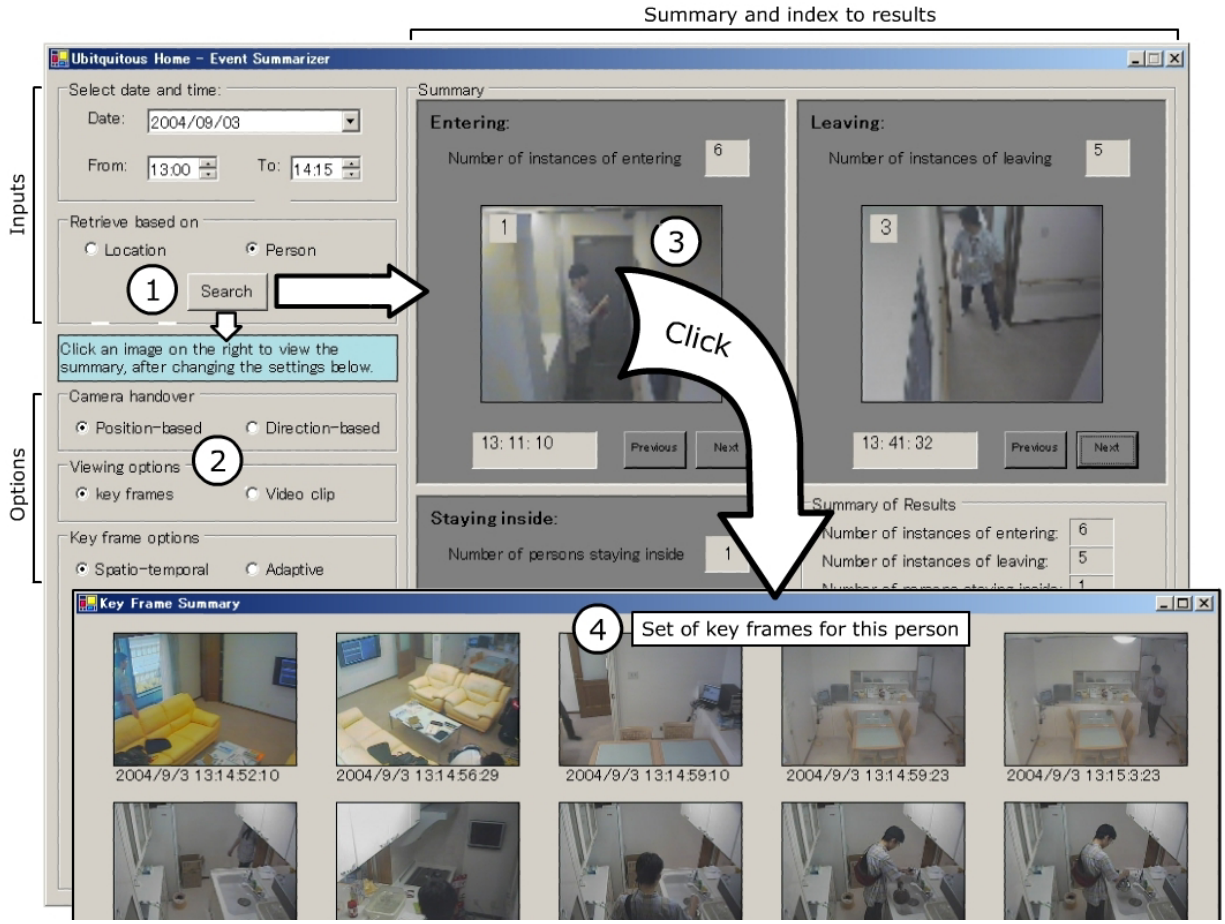

Figure 4. User interaction with system.

we lose information in case of fast movement or larger time interval. Another approach is *spatial sampling*, where key frames are sampled according to the motion of the person in the environment. In case of this work, we implement spatial sampling by extracting a key frame whenever the camera that is used to show the person is changed. A key frame summary created in this method can help tracing the path the person took. The weakness in this method is not being able to extract key frames for actions performed when the person is in the same view.

A combination of spatial and temporal sampling can improve the results of key frame extraction as the methods can 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 same. We implement an *adaptive spatio-temporal* sampling algorithm based on this hypothesis.

Table 2 summarizes the algorithms we designed for key frame extraction. In all entries, T is a constant time interval.

3.4 Retrieval of Summarized Data

The results are stored in a database to be queried through a graphical user interface. A query is initiated by entering the time interval for which the summary is required. For the 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

Table 2. Algorithms for key frame extraction

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

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 4 is a screenshot of the system.

4. EVALUATION EXPERIMENT

We decided to evaluate the algorithms we implemented for key frame extraction, with the following objectives:

(1) Evaluation of the algorithms we designed for key frame extraction to select the best algorithm and the correct value for the parameter T

(2) Investigate the possibility of extracting an average set of key frames based on those selected by a number of persons

(3) If such a set can be obtained, use it for defining accuracy measures for the extracted key frame sequences

(4) Use the average key frame sets as targets for improving the algorithms or designing new algorithms

(5) Obtain feedback on the performance of the existing algorithms for key frame extraction and identify requirements for better performance.

Since it was not possible to find an existing method of evaluation available to fulfill the above, we decided to design and conduct our own evaluation experiment. The design of the experiment was independent of the way the video has been created, making it usable for evaluation of any key frame extraction algorithm in general. The experiment consists of a key frame extraction task, comparison of key frames, and providing comments and suggestions. The following sections describe the experiment in detail.

4.1 Key Frame Extraction Task

The key frame extraction task is based on a video sequence created by video handover, hereafter referred to as a *sequence*. The task consists of three sections, as described by the following paragraphs.

In the first section, the test subject browses the sequence, and selects key frames to summarize the sequence based on their own choice. There is no limit in terms of either the time consumed for selection or the number of frames selected. This section of the experiment is performed first in order to ensure that seeing the key frames extracted by the system does not influence the subjects.

Table 3. Criteria for evaluating individual frame sets

Criterion	Responses
1. Number of key frames as compared to the duration of the sequence	(a) Too few (b) Fine (c) Too many
2. Percentage of redundant frames	(a) None (b) Less than 25% (c) 25%-50% (d) More than 50%
3. Number of important frames missed	(a) None (b) 1 to 5 (c) 6 to 10 (d) More than 10

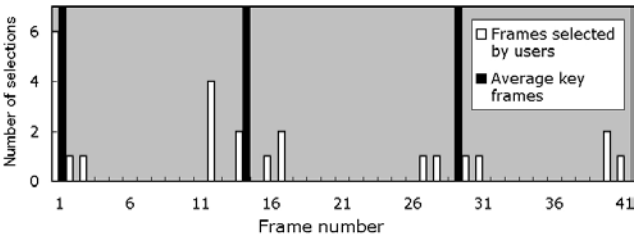

Figure 5. Average key frames.

In the second section, the subject evaluates sets of key frames (hereafter referred to as *frame sets*) corresponding to the same sequence, created automatically by the system using different algorithms. A total of seven frame sets are presented for each sequence; one created by spatial sampling, two each for the other algorithms with $T = 15$ s and 30 s. These were presented to the subject in a random order, to ensure that the evaluation is not affected by the order of presenting the results. The subjects rank each frame set against the criteria presented in Table 3.

In the third section, the subject compares different frame sets and selects the frame set that summarized the sequence best. For the frame set they selected, they answer the following questions:

(a) Why do you find it better than other sequences?

(b) In what ways can it be improved?

4.2 Experimental Procedure

Eight voluntary subjects took part in the experiment. None was involved with the design of algorithms for key frame extraction. Each subject was briefed about the task at the beginning of the experiment and written instructions were provided. One of the authors was available throughout the experiment to provide additional clarifications if needed.

Each subject completed four repetitions of the key frame extraction task, on four different sequences. The sequences consisted of a combination of attributes such as the length, the actions the persons in sequences performed, interaction with objects, etc. The subjects were allowed to watch the sequences as many times as they desired. Breaks were allowed between repetitions. The subject concludes the task by stating additional comments and suggestions, if any.

The subjects took 65 to 120 minutes to complete the experiment. This time included short breaks between repetitions.

5. RESULTS

5.1 Average Key Frame Selection

The key frame sets selected by different subjects had different numbers of key frames. However, visual inspection showed that there are a considerable proportion of common key frames. Figure 5 presents a histogram of key frames selected by the subjects, $f(n)$ for a portion of one sequence. It is evident that key

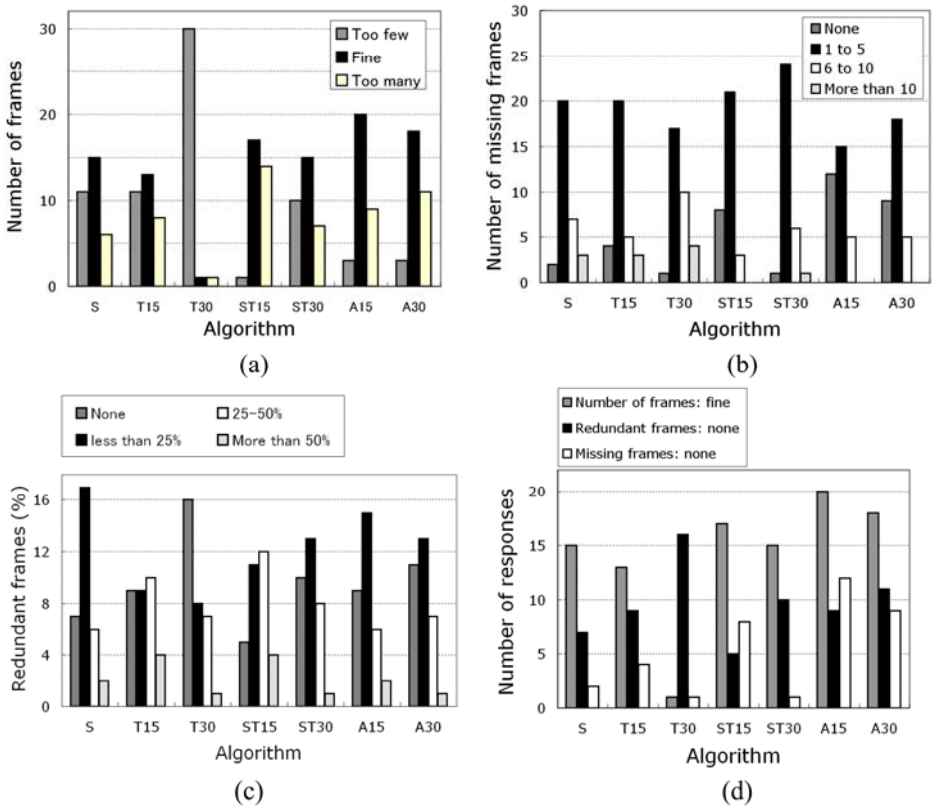

Figure 6. Comparison of votes for the responses.

frames selected by different subjects form small clusters corresponding to actions and events they wished to include in their summaries.

The following algorithm was used to form an *average key frame set* for each sequence. First, we examine $f(n)$ from $n = 0$ and identify non-overlapping windows of 10 frames, within which 50% or more of the subjects selected a key frame. From each window W, an average key frame k is extracted using the following equation:

$$k = \left[\frac{\sum_{n \in W} n f(n)}{\sum_{n \in W} n} \right]$$

Table 4. Comparison of the number of key frames

Sequence Number	1	2	3	4
Average value of the number of key frames selected by subjects	6.5	8	13	32.8
Number of key frames in the average key frame set	6	6	11	30

The average key frames for the frames corresponding to Figure 5 are indicated by black markers on the same graph.

Table 4 presents a comparison of the average number of key frames the users selected and the number of key frames in the average key frame sets. The numbers are nearly equal. This is not possible unless there is a strong agreement on the actions and events to be selected as key frames, among different subjects. Therefore, we suggest that it is possible to use these key frame sets in place of ground truth for evaluation of the algorithms for key frame extraction. Furthermore, we propose that the algorithms can be improved by modifying them to retrieve key frame sequences that are closer to the average key frame sets.

Table 5. Abbreviations for labeling frame sets

Abbreviation	Description
S	Spatial sampling
T15	Temporal sampling with $T = 15$ s
T30	Temporal sampling with $T = 30$ s
ST15	Spatio-temporal sampling with $T = 15$ s
ST30	Spatio-temporal sampling with $T = 30$ s
A15	Adaptive spatio-temporal sampling with $T = 15$ s
A30	Adaptive spatio-temporal sampling with $T = 30$ s

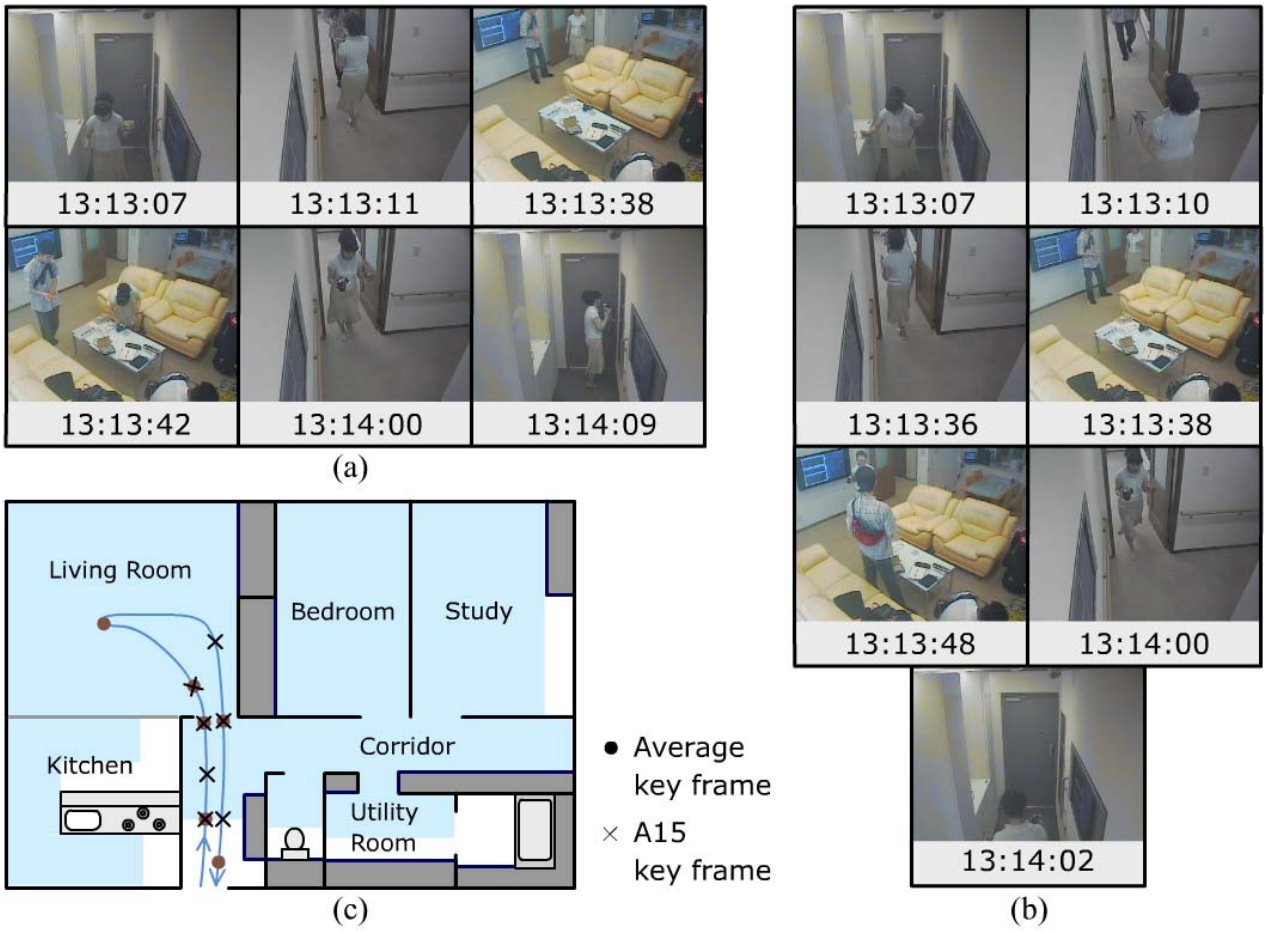

Figure 8. Comparison of average and A15 key frames.

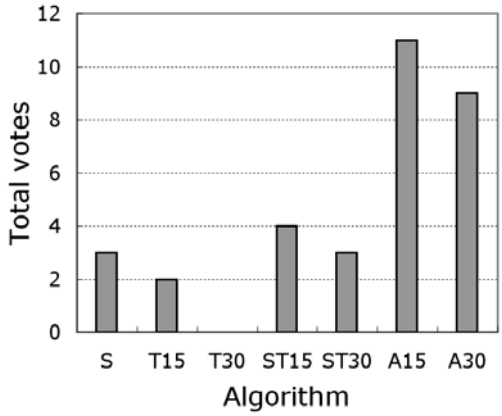

Figure 7. Comparison of votes for the best responses.

Figure 6 (a), (b) and (c) compares the responses from the test subjects for each criterion stated in Table 3. The abbreviations used to denote the algorithms are explained in Table 5. The responses for T30 in Figure 6(a) suggest that 30 seconds is too large an interval between key frames for video captured in this environment. However, the number of redundant frames or that of missing frames cannot be considered alone to select the best method, since these two measures are somewhat analogous to the *precision* and *recall* measures of information retrieval. Therefore, the best category of responses for each criterion was compared to find out which algorithm has the best overall performance (Figure 6d). It is evident that adaptive sampling has performed much better than the other algorithms. The method A15 was found to perform best in terms of the number of frames and not missing frames. The method A30 performs slightly better in terms of less redundant frames, compared with method A15. The sum of responses for the three categories is higher for the method A15, suggesting that $T = 15$ s is more suitable.

Figure 7 presents the votes received by each method for the best frame set. The results are consistent with those from the previous section of the evaluation. The methods A15 and A30 acquired 62% of the total votes, indicating that adaptive spatio-temporal sampling performs far better than the other algorithms and 15 s is a more suitable value for the parameter T .

5.3 Comparison with average key frames

The frame sets were compared with the corresponding average key frame sets subjectively. It was observed that the key frames extracted using A15 are the most similar to the average frames. Figures 8 (a) and 8 (b) show the average key frames and the frame set created by this method respectively, for one sequence. Figure

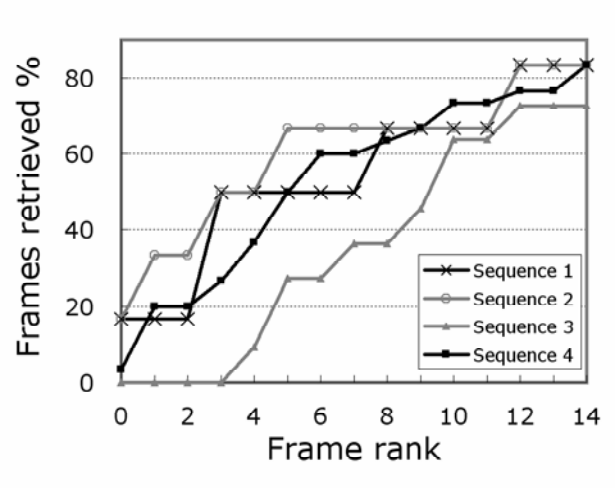

Figure 9. Cumulative performance of key frame extraction.

8(c) shows the path of the person in the sequence, with locations of the person when the key frames were sampled. The algorithm failed to capture the key frame corresponding to the girl picking a camera from the stool. It extracted two redundant frames as she was within the same view for a longer time.

To evaluate the performance of this key frame extraction method quantitatively, we define the rank n performance, R_n of the method as:

$$R_n = \frac{K_n}{N} \times 100\%$$

where,

K_n = number of occasions a key frame is present within n frames from that of the average key frame set

N = number of frames in the average key frame set

Figure 9 plots the cumulative performances against n. The results show that it is possible to extract key frames within a difference of 3 s, with an upper bound of around 80%, using only floor sensor data with this method.

5.4 Descriptive Feedback

The questionnaire included two qualitative questions about the frame set that the subject rated as the best. Answers to the first question "Why do you find it a better summary than other sequences?" are listed below (number of occurrences of each response is indicated in parentheses):

- Minimum number of key frames missed (11)
- Minimum number of redundant frames (6)
- Right number of key frames (5)
- Complete summary (3)
- Match well with own selection (2)
- Full view of person in most of the key frames (2)

Answers to the second question "In what ways can it be improved?" included:

- Add key frames to show interaction with other persons and objects (4)
- Remove redundant key frames (2)
- Try to get a full view of the person in a key frame (2)
- Add key frames to show corners in walking path (1)

Most of the subjects considered it important not to miss any important key frames when summarizing a video, in agreement with the results form the previous section of the experiment. The comments demonstrate that the test subjects desire the inclusion of key frames corresponding to human object and human-human interaction to be included in an improved set of key frames. This was consistent with the observation that such key frames were included in the average key frame sets. The results were not significantly different for sequences with different durations or actions. The only exception was low performance with sequence 3 as shown in Figure 9. This was mainly due to the fact that the person shown in this sequence moves slower and stops for some time in a number of places. Therefore the picked up frames can be a bit further from what the algorithm sampled, but still they show the same event or action.

6. DISCUSSION

It is evident that the difference of performance between the two adaptive methods is very small. The reason for this is that the extraction depends on the behavior of the persons in the video sequence, rather than the value of T . Both algorithms can produce the same result in some situations; for example, if a person walks in a way that the view changes every 5 seconds.

The technique used to construct average frame sequences currently considers only the difference in time. For parts of the video with little or no motion, the users may pickup key frames for the same action within a larger gap than 10 frames. Considering the pixel-wise differences between images may be useful to achieve better results in such cases.

Some of the subjects commented that automatic annotations to key frames are desirable. However, annotations will be useful only if they are at a higher semantic level. For example, "entered the house" is not a useful annotation, as this can be understood easily by observing the frame. Image analysis on the key frames and obtaining supplementary data from additional sensors can be helpful in annotation at a higher level.

The evaluation for key frame extraction is not specific for those created by footstep segmentation and video handover. With minor changes where necessary, the evaluation can be applied to summarization of arbitrary types of video. However, the main problem in using this technique is the large amount of time consumed for manual key frame extraction. One way to solve this problem is to present the video in terms of an initial set of key frames with high redundancy and let the user summarize it in a hierarchical fashion.

Most of the subjects desired to extract key frames showing a full view of the person where possible. This suggests that better summaries can be realized if the handover can maximize the

availability of a full view after a shot boundary. Furthermore, occlusion by other persons in the environment should be considered while selecting the view for the key frame extraction.

The floor sensors facilitate tracking people with less computational effort compared to using image analysis. However, they are not applicable to any environment. Moreover, movement of furniture can cause noise and clutter, making tracking difficult. The accuracy of using RFID tags together with floor sensors is now under investigation.

7. CONCLUSION

We have implemented personalized video summarization for a ubiquitous environment with a large number of cameras, by analyzing signals from pressure based sensors mounted on the floor. A number of algorithms for extracting key frames from the video data were implemented. An experiment was designed and conducted for evaluating the performance of these algorithms, selecting the most suitable algorithm and identifying ways to improve key frame extraction. The experiment can be applied to evaluate key frame extraction algorithms on any type of video.

It was observed that there is strong agreement among different persons on selection of key frames. An algorithm that is adaptive to the rate of the footsteps of the person was found to extract key frame sequences that are the best in terms of the number of redundant and missing key frames. Quantitative evaluation based on average sets of key frames showed that about 80% of the most desired key frames can be retrieved using the current algorithms that analyze only the footsteps of a person.

Future work will focus on extracting key frames for interaction among persons and between a person and an object. Automated generation of camera mappings will make camera handover easily adaptable to different settings. Detection of higher-level features such as actions can greatly enhance the key frame sets. The average key frame sets can be used as targets for improving algorithms for key frame extraction, before they are tested using a complete experiment. An interesting future direction is to investigate the possibilities of presenting audio data together with a key frame summary.

ACKNOWLEDGMENTS

We thank Mak Mei Poh for helpful discussions on the design of experiments. This work is partially supported by the Keihanna Human Info-communications Research Center of the National Institute of Information and Communications Technology (NICT), Japan.

REFERENCES

[1] T. Yamazaki, “Ubiquitous Home: Real-life Testbed for Home Context-Aware Service”, In *Proceedings of Tridentcom2005*, pp.54-59, February 23, 2005.

[2] N. Sebe, M. S. Lew, X. Zhou, T. S. Huang, E. Bakker, “The State of the Art in Image and Video Retrieval”, International Conf. on Image and Video Retrieval (CIVR'03). (2003) 1--8

[3] J. R. Wang, N. Prameswaran, X. Yu, C. Xu, Qi Tian: Archiving Tennis Video Clips Based on Tactics Information, In proc. of the 5th Pacific Rim Conf. on Multimedia. (1996)

[4] A. Haubold, J. R. Kender, “Segmentation, Indexing, and Visualization of Extended Instructional Videos”, *CoRR* cs.IR/0302023 (2003)

[5] A. Divakaran, I. Otsuka, R. Radhakrishnan, K. Nakane, M. Ogawa: Audio-Assisted Video Browsing for DVD Recorders. PCM (2) 2004: 27-33

[6] K. Morisawa, N. Nitta, Noboru Babaguchi: Video Scene Retrieval with Sign Sequence Matching Based on Audio Features. PCM (2) 2004: 121-129

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

[8] K. Aizawa, S. Kawasaki, T. Ishikawa, T. Yamasaki, “Capture and retrieval of life log”, Proc. ICAT2004, pp.49-55, Nov.30-Dec.2, 2004.

[9] 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.

[10] G. A. Abowd, I. Bobick, I. Essa, E. Mynatt, and W. Rogers: The Aware Home: Developing Technologies for Successful Aging, In proc. AAAI Conf. 2002, Canada, July 2002.

[11] R. J. Orr and G. D. Abowd: The Smart Floor: A Mechanism for Natural User Identification and Tracking, In Proc. of the 2000 Conf. on Human Factors in Computing Systems.

[12] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata: Memory Cues for Meeting Video Retrieval, Proc. CARPE 2004.

[13] T. Mori, H. Noguchi, A. Takada, T. Sato, “Sensing Room: Distributed Sensor Environment for Measurement of Human Daily Behavior”, First International Workshop on Networked Sensing Systems (INSS2004), pp.40-43, 6.

[14] K. Matsuoka and K. Fukushima, “Understanding of Living Activity in a House for Real-time Life Support”, Proc. SCIS & ISIS 2004, pp.1-6, Japan (2004)

[15] L. Liu and G. Fan, "Combined Key-frame Extraction and Object-based Video Segmentation", *IEEE Trans. Circuits and System for Video Technology*, July, 2004.

[16] M. R. Naphade, J. R. Smith, “On the Detection of Semantic Concepts at TRECVID”, Proc. ACM Multimedia 2004.

[17] TRECVID 2005 Guidelines, http://www-nlpir.nist.gov/projects/ tv2005/tv2005.html, National Institute of Standards and Technology, USA, 2005.

[18] S. Kawasaki, T. Ishikawa, T. Yamasaki, K. Aizawa, “Effective Life-Log Video Summarization Based on Sampling of Sensor Data”, Proc. IEICE MVE, March 2005.

[19] Song, X., Fan, G., Joint Key-Frame Extraction and Object-Based Video Segmentation, Motion05 (II: 126-131).

[20] G. C. de Silva, T. yamasaki, T. Ishikawa, K. Aizawa, "Video Handover for Retrieval in a Ubiquitous Environment Using Floor Sensor Data", In proc. ICME 2005, July 2005.

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