学位論文

Multimedia Experience Retrieval in a Ubiquitous Home
(ユビキタスホームにおける体験情報処理と検索)

2006年12月14日
指導教員　相澤　清晴　教授

東京大学　新領域創成科学研究科
基盤情報学専攻

De Silva Gamhewage Chaminda
デシルヴァ　ガムヘワゲ　チャミンダ

Acknowledgment

First of all, I would like to thank my supervisor, Prof. Kiyoharu Aizawa, for his excellent supervision and guidance during the past three years. His devotion to my development was more than what can be expected from a supervisor. Dr. Toshihiko Yamasaki was always helpful despite his busy schedule, and ensured that I could carry out my research smoothly. I would like to express my gratitude to my former supervisors, Dr. Michael J. Lyons and Dr. Liyanage C. de Silva, for their continuing support for my progress. Many thanks are due to the members of Aizawa Laboratory, for participating in the tedious experiments and providing valuable feedback.

The staff at NICT Keihanna Info-communication Center provided tremendous support in conducting experiments and collecting data. Ms. Kaori Ono, Ms. Chiho Miyao, Ms. Hiromi Yamazaki and the other administrative staff in the university were always ready to help. My special thanks go to Ms. Fusako Ide and the staff of the International Liaisons office, for their efforts in improving my Japanese language skills, and helping with the (sometimes scary) formalities.

Many thanks are due to those people who were close to me during my study in Tokyo University. Iguchi-san and Keiko-san, my partner family from Mitsui Volunteer Network, were actually family to me with their love, kindness and concern. I am thankful to my good friends; Pamela, Lai (Ivan), Ahmet, Claus, Sudanthi, and many others. They were always around; sharing their thoughts and expertise with me, helping me to get over the hard times, and spending free time together with me. They made my stay here a memorable one.

Last but not least, there are a few people without whom this thesis could have been impossible. I thank my parents and family for their constant love and support. I am in debt to Yuki-san, for filling loads of application forms at a time I did not know how to read or write Japanese, to ensure that I get admission to Tokyo University. Finally, I dedicate this thesis to Mei, for her persistent support which ensured that I complete it.

Table of Contents

List of figures

List of Tables

Abstract

Automated capture and retrieval of multimedia experiences at home is interesting due to the wide variety and personal significance of such experiences. However, this is a difficult task with several challenges in different aspects. The number of sensors required for complete capture of experiences is quite large. Continuous capture from such a collection of sensors results in a large amount of multimedia content that is much less structured compared to those from any other environment. Experiences are difficult to recognize by automated analysis of sensor data, due to their high semantic level. Queries for retrieval will be at different levels of granularity, calling for well designed user interaction.

In this research, we focus on capturing and retrieval of personal experiences in a ubiquitous environment that simulates a home, with the objective of creating a multimedia chronicle that enables the residents to retrieve the captured media using simple, interactive queries. A large number of cameras and microphones continuously record video and audio at desired areas of the house. Pressure based sensors, mounted on the house floor, record context data corresponding to the footsteps of residents.

Our approach to achieve efficient multimedia retrieval from this large collection of data is based on adaptive source selection using both context and content analysis. Data from floor sensors are analyzed to segment footstep sequences of different persons, which are then used for the creation of video clips while automatically changing cameras and microphones to keep the person in view and hear the sounds in his/her surroundings. These videos are further summarized into sets of key frames, allowing the

users to view a compact and complete summary of their content. Audio data from the microphones are segmented and classified into different categories of sounds, to retrieve the sounds and video showing the locations where the sounds are heard. Basic analysis of image data facilitates the detection of selected events that take place inside the house. Floor sensor data are analyzed in combination with other sensory modalities, for recognition of some common actions inside the house. The results are written to a central a relational database, where they can be fused for accurate detection of activities. The users, who also are the residents, retrieve their experiences from the database through a graphical user interface by submitting interactive queries. This interface was designed based on the concepts of hierarchical media segmentation and Interactive retrieval, to facilitate effective retrieval with a small amount of manual data input using only a pointing device. Visualizations of different types of data at various levels of detail were included to help the user to retrieve required media and understand the results.

Each functional component of the system was evaluated individually, to ensure that it provides accurate results to the user and the other components using the results. We used standard accuracy measures and experiments where available, while designing experiments and defining new accuracy measures where necessary. We conducted a user study for the purposes of gathering system requirements and evaluating the overall system. A family who actually lived in ubiquitous home was selected as the subjects for this study.

Hierarchical clustering of floor sensor data followed by media handover enabled the creation of personalized video clips using a large number of cameras, with a reasonably good audio quality. An adaptive algorithm enabled retrieval of more than

80% of the key frames required for a complete summary of the video. Silence elimination and false positive removal from audio data produced results with a high accuracy of 98%. The scaled template matching algorithm we propose is able to localize sound sources with an average accuracy of 90%, despite the absence of microphone arrays or a beam-forming setup. Accuracy of audio classification using only time domain features is above 83%. Basic image analysis facilitated detection of events that are useful in understanding the activities that take place inside the house. Action detection using multiple sensory modalities yielded an average accuracy of approximately 78%.

The residents who evaluated the system found it useful, and enjoyed using it. They found the system easy to learn and usable. The requirements they identified and the feedback they provided were valuable in improving the system.

List of Publications

Journal Articles

1. G. C. de Silva, T. yamasaki, K. Aizawa, "Sound Source Localization for Multimedia Retrieval in a Ubiquitous Environment", *IPSJ Letters on Information Science and Technology (情報科学技術レターズ), Special Issue for FIT 2006*, pp. 197-199.

2. G. C. de Silva, T. yamasaki, K. Aizawa, "An Interactive Multimedia Diary for the Home", Submitted to *IEEE Computer Magazine, Special Issue on Human Centered Multimedia, April 2007.*

3. G. C. de Silva, T. yamasaki, K. Aizawa, "Sound Source Localization Based on Energy Distribution Template Matching for a Ubiquitous Environment", Submitted to *IEEE Transactions in Multimedia.*

4. G. C. de Silva, T. yamasaki, K. Aizawa, "Ubiquitous Home: Retrieval of Experiences in a Home Environment", Submitted to *Transactions of IEICE.*

Reviewed Conference Papers

1. G. C. de Silva, T. Yamasaki, K. Aizawa, "Interactive Experience Retrieval for a Ubiquitous Home" *Proceedings of ACM workshop on Capture, Archival and Retrieval of Personal Experiences (CARPE) 2006*, pp.45-48.

2. G. C. de Silva, T. Yamasaki, K. Aizawa, "Creation of an Electronic Chronicle for a Ubiquitous Home: Sensing, Analysis and Evaluation", *Proc. IEEE Workshop on Electronic Chronicles 2006.* pp.70-78

3. G. C. de Silva, T. Yamasaki, K.Aizawa, "Person Tracking and Multi-camera Video Retrieval Using Floor Sensors in a Ubiquitous Environment" *Proceedings of Pacific-rim Conference in Multimedia (PCM) 2005*, pp. 1005-1016.

4. G. C. de Silva, B. Oh, T. Yamasaki, K. Aizawa, "Experience Retrieval in a Ubiquitous Home" *Proceedings of ACM workshop on Capture, Archival and Retrieval of Personal Experiences (CARPE) 2005.* pp. 35-44

5. G. C. de Silva, T. Yamasaki, K. Aizawa, "Evaluation of Video Summarization for a Large Number of Cameras in Ubiquitous Home", full paper accompanied by video figure, *Proc. ACM Multimedia 2005.* pp.820-828

6. G. C. de Silva, T. Ishikawa, T., Yamasaki, K. Aizawa, "Person Tracking and Multi-camera Video Retrieval Using Floor Sensors in a Ubiquitous Environment", *Proceedings of International Conference in Image and Video Retrieval (CIVR) 2005*, pp. 297-306

7. Gamhewage C. de Silva, T. yamasaki, T. Ishikawa, K. Aizawa, "Video Handover for Retrieval in a Ubiquitous Environment Using Floor Sensor Data", In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME) 2005.*

Non-reviewed Conference Papers

1. G. C. de Silva, T. Ishikawa, T. Yamasaki, Kiyoharu Aizawa, "Audio Segmentation Using a Large Number of Microphones for Multimedia Retrieval in a Ubiquitous Environment", *Proceedings of IEICE National Conference 2006*, p. 276.

2. G. C. de Silva, T. yamasaki, K. Aizawa, "Selection from a Large Number of Audio and Video Sources for Personalized Video Retrieval in a Ubiquitous Environment", *Proceedings of FIT 2005*, pp. 267-268.

3. G. C. de Silva, T. yamasaki, K. Aizawa, "Video Summarization for a Large Number of Cameras Using Floor Sensors in a Ubiquitous Environment", *Proceedings of ITE National Conference 2005*, M-025.

4. G. C. de Silva, T. Ishikawa, T. Yamasaki, K. Aizawa, "Video Retrieval in a Ubiquitous Environment with Floor Sensors", *Proceedings of IEICE National Conference 2005*, p. 165.

Chapter 1

Introduction

Humans have always had the tendency to record their experiences using some means, even before the earliest civilization. The earliest examples for such records are ancient cave paintings that date thousands of years back. With the advancement of technology, more and more methods for this task became both available and affordable. As a result, the size of the content recorded from one's life has greatly increased over the past few decades. In parallel to this, there has been a growing interest in research related to continuous capture and retrieval of personal experiences.

1.1. Retrieval of experiences from life at home

Automated capture of experiences taking place at home is interesting owing to 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 humans do not want to keep themselves out of the experience to 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. Content captured at home 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?"

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.

1.2. Motivation

Investigation in to automated retrieval of experiences at home can be useful in several other aspects, in addition to the significances mentioned above. This topic encompasses the general research areas of multimedia retrieval and ubiquitous environments. However, a home is much less controlled compared to the other ubiquitous environments used in related research. Video captured at home are unstructured content, marking a significant contrast from news, sports or instructional video which are the common inputs for automated retrieval. Therefore, the selected topic will pose several research challenges, with prospects of significant contributions to these areas. The outcomes of this research will be applicable in areas with practical significance, such as automated surveillance, elder care, and automated video summarization.

1.3. Objectives

This thesis presents our work on capturing and retrieval of personal experiences in a ubiquitous environment that simulates a home. The primary objective is to create an *electronic chronicle* [1] that enables the residents of the house to retrieve the captured video using simple, interactive queries within a short search time. To achieve this, we design and implement algorithms that perform unsupervised data mining algorithms on

context data from pressure based sensors mounted on the house floor. Audio analysis, segmentation and classification are used to both complement context based retrieval and achieve content based retrieval. Activity detection is facilitated by combining the results of video, audio and floor sensor data. Accuracy measures are defined and experiments 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.

1.4. Organization of the thesis

The remainder of this thesis is organized as follows: Chapter 2 outlines recent related research; Chapter 3 described Ubiquitous Home, the environment where we capture data for this work. An overview of the system is presented in Chapter 4. Chapters 5, 6 and 7 describe the algorithms used to analyze data from different types of sensors for retrieval of multimedia experiences. Chapter 8 describes the design of user interaction with system. The user study conducted for evaluating the overall system is described and the results presented in Chapter 9. Chapter 10 concludes the thesis, suggesting possible future directions.

Chapter 2

State of the Art

This research combines the work from the research areas of *Ubiquitous Environments*, *Multimedia Retrieval*, and applies them to form a system capable of *capturing and retrieving personal experiences*. The following sections of this chapter present the state of the art of these research areas.

2.1 Ubiquitous Environments

Ubiquitous environments are equipped with a large number of sensors of different types, enabling acquisition of data regarding the events that take place within 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 three major categories. The first category aims at providing services to the people in the environment by detecting and recognizing their actions. Such environments serve as *information appliances*; examples are numerous *Smart Home* projects that intend to make daily life comfortable [2] [3], and the *Aware Home Project* [4] for supporting elderly residents. Basic activities such as opening and closing of doors can be recorded using switch-based sensors [5]. Numerous types of sensors are used for tracking and detection of the persons and recognize their activities. Use of cameras and image analysis for this purpose is common. In Easy Living Project [6][7] and Intelligent Space [8], the positions of humans are detected using multiple cameras. However, alternative

methods such as Radio Frequency Identification (RFID) tags [9][10], optical tags [11] and Infra-red based motion sensors [12] have been used where image acquisition and analysis is not possible due to issues such as privacy, disk space, and computational cost.

The second category of ubiquitous environments aims at storing and retrieval of media captured within the environments, 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. Applications in this category include meeting video retrieval [13][14] and summarization of instructional video[15][16]. Some of the projects, such as *CHIL* [17], attempt to combine both the above directions by supporting user interaction real-time and using retrieval for long term support.

The third category is surveillance, where the data captured in the environment are processed to obtain information that help to raise alarms, in order to protect the environment and people who use it. Video is highly prospective as an input modality for this purpose, due to its non-intrusive nature and rich information content. Research on automated video surveillance has been growing rapidly during the past few years. A recent review of the state of the art is found in [18]. Systems based on single or multiple cameras, both stationary and moving, have been designed an implemented for automatic detection, tracking and recognition of humans and there actions [19][20][21][22][23]. Some of these researches try to combine data from other sensors, to improve accuracy [24][25][26][27]. However, at the current state, none of these systems have sufficient accuracy to be deployed in practical situations for fully automated surveillance.

Therefore, some of the recent researches focus on assisting humans monitoring the environment rather than fully automated surveillance [28].

2.2 Multimedia Retrieval

A detailed discussion of the state of the art of multimedia retrieval can be found in [29], while a more recent review is available in [30]. Most of the existing researches deal with previously edited single stream broadcast video with specific content [31][32][33]. Example applications include news video retrieval [34][35][36], and sports video summarization and indexing [37][38][39]. For such data, the common approach is content analysis making use of domain knowledge where applicable [40]. However, the use of context data where available can improve the performance greatly [41].

2.3 Multimedia Retrieval for Ubiquitous Environments

There are several ongoing projects that work on multimedia retrieval for ubiquitous environments. The *Ubiquitous Sensor Room* [42] 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. [43] utilize graphical representations of important memory cues for interactive video retrieval from a ubiquitous environment. The *Sensing Room* [44] 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. [45] 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.

2.4 Capture and Retrieval of Personal Experiences

The research theme of capture and archival of personal experience is quite new, although the emergence of such research had been predicted much earlier in science literature [46]. There have been a few researches on capturing of life media using wearable cameras during the last decade [47][48][49]. Recent research initiatives such as *The Microsoft Memex Project* [50] have prompted a growth in this area, during the last couple of years. The main difference in this theme from the other work on multimedia retrieval is the personal nature of data and the high semantic level of the experiences retrieved. The researches in this area capture data from wearable, pervasive and other types of sensors over a long period of time and then analyze the data to for classification of actions, events and experiences [51]. *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 [52]. The *MyLifeBits* system collects data about a person's usage of computers, documents and television, and attempt to organize these data in a manner that allows faster retrieval [53].

2.5 Summary

While there has been a considerable amount of research in the individual areas of multimedia retrieval and ubiquitous environments, research combining these two areas has been relatively new and limited to applications with either manual monitoring or

relatively short periods of data acquisition. The selected topic of Multimedia experience retrieval from a home like ubiquitous environment is both novel and challenging. The outcomes of such research will contribute to the progress of both areas of research, facilitating efficient use of hardware and media capture technologies.

Chapter 3

Ubiquitous Home

The primary requirement for this research is a home-like ubiquitous environment that is equipped with a sufficient number of cameras and microphones in order to capture the media that the residents would like to retrieve, and able to capture media for a long period of time. We selected the *Ubiquitous Home* [54], built in the Keihanna Human Info-communication Laboratory of the National Institute of Information and Communication Technology of Japan, as the environment for this work. Simulating a two-room house, it has been has been designed to provide a testing ground for ubiquitous sensing in a household environment. The following sub-sections describe the sensor arrangement, data collection and main issues concerning capture and retrieval.

3.1 Sensors and Data Acquisition

Figure 1 shows the floor plan and the sensor layout of the ubiquitous home. The non-private areas of the house are 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 5 frames per second and stored in JPEG file format. The frame rate is low due to storage space restrictions, but this frame rate is adequate given the pace of human behavior in a household environment. 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.

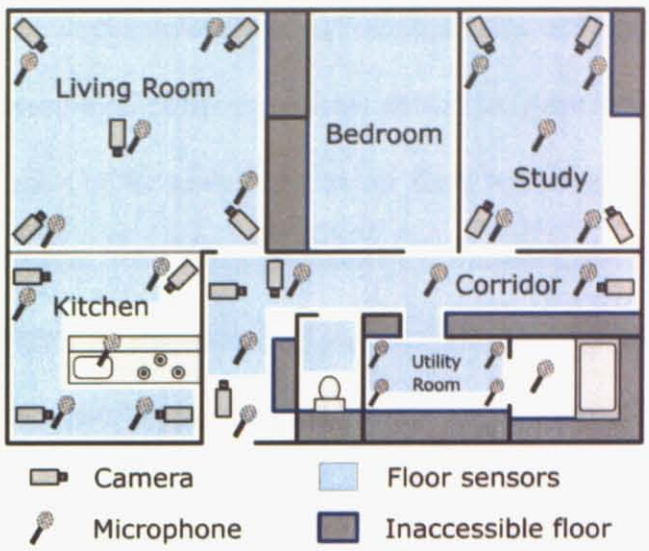

Figure 1: Ubiquitous home sensor layout.

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 increases and crosses a specific threshold, it is considered to change its state to '1'. The state is reset to '0' when the pressure becomes lower than the threshold again. Each state transition is recorded in a database with the timestamp, coordinates of the sensor, and the new state of the sensor.

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 real-life situations properly, a series of "real-life experiments" were conducted. In each experiment, a family lived in Ubiquitous home for a period of 1-2 weeks. The families lead their normal lives during this stay. They were not restricted in terms of the amount of time that they spent in the house. The family members went to work/school during weekdays; they cooked and had meals in the house; there were occasional visitors; and everybody went out at times. Families with members of different ages participated in different experiments.

No manual monitoring of data was done during the experiments after adjustments before the experiments. The images, audio and sensor data were stored separately with timestamps for synchronization. The processing was performed offline. However, the algorithms were designed so that they can be adapted for real time processing.

3.3 Issues Related to Capture

The main issue in capturing data in ubiquitous home is the large amount of disk space required. Each day of continuous capture results in consumes about 500GB of disk

space. The current storage capacity of ubiquitous home allows only 14 days of continuous data acquisition, thereby limiting the capability of acquiring long term behavioral patterns.

The high consumption of disk space is partially due to low compression and disk fragmentation, resulting from storing a large number of small files. For instance, the size on disk for video data captured during a single day is about 420GB, while the actual total file size is only about 220 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.

A few issues arise from the construction, installation and interfacing of floor 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.

One day of continuous capture in ubiquitous home results in 408 hours of video and 600 hours of audio. This long duration of the content makes automated retrieval essential for efficient experience retrieval from this environment. The following chapters outline the system that we propose for this purpose and describe the algorithms that are used for multimedia retrieval using different types of sensory data.

3.4 Discussion

It is evident that the ubiquitous home can be made more functional in terms of capturing daily life, by installing additional sensors of different types. Infra-red based motion sensors can be used in combination with floor sensors, for accurate motion tracking. Sensors indicating the opening and closing of doors can provide highly accurate data. Other sensory modalities, such as temperature and light level, can be measured with simple sensors and recorded at the expense of small amount of disk space. These sensors can be connected using a wireless network, making installation easier. However, it should be noted that we were not in control of deciding the sensor arrangements of ubiquitous home. Therefore, we decided to work with existing sensor data for this research.

Chapter 4

System Overview

4.1 Issues

The main problem in multimedia retrieval from ubiquitous home is caused by the large number of sources and the huge amount of data. An approach based on exhaustive content analysis will be computationally very expensive. 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 select sources that convey the most amount of information based on context data. Only the selected sources are queried to retrieve data and these data are analyzed further for retrieval, thereby minimizing the computational effort on content analysis. However, at the same time, the redundancy caused by the presence of a large number of sensors is utilized to improve the accuracy of retrieval.

4.2 Outline of the Proposed System

Figure 2 is a functional block diagram of the system that we propose for efficient multimedia experience retrieval from ubiquitous Home. Data from floor sensors are analyzed for retrieving footstep sequences, video clips and key frames. Audio data from the microphones are segmented and classified into different categories of sounds, to retrieve the sounds and video showing the locations where the sounds are heard. Analysis of image and floor sensor data facilitates the detection of some events that take place inside the house. The results are written to a central a relational database, where they can be fused for accurate detection of activities. The users, who also are the

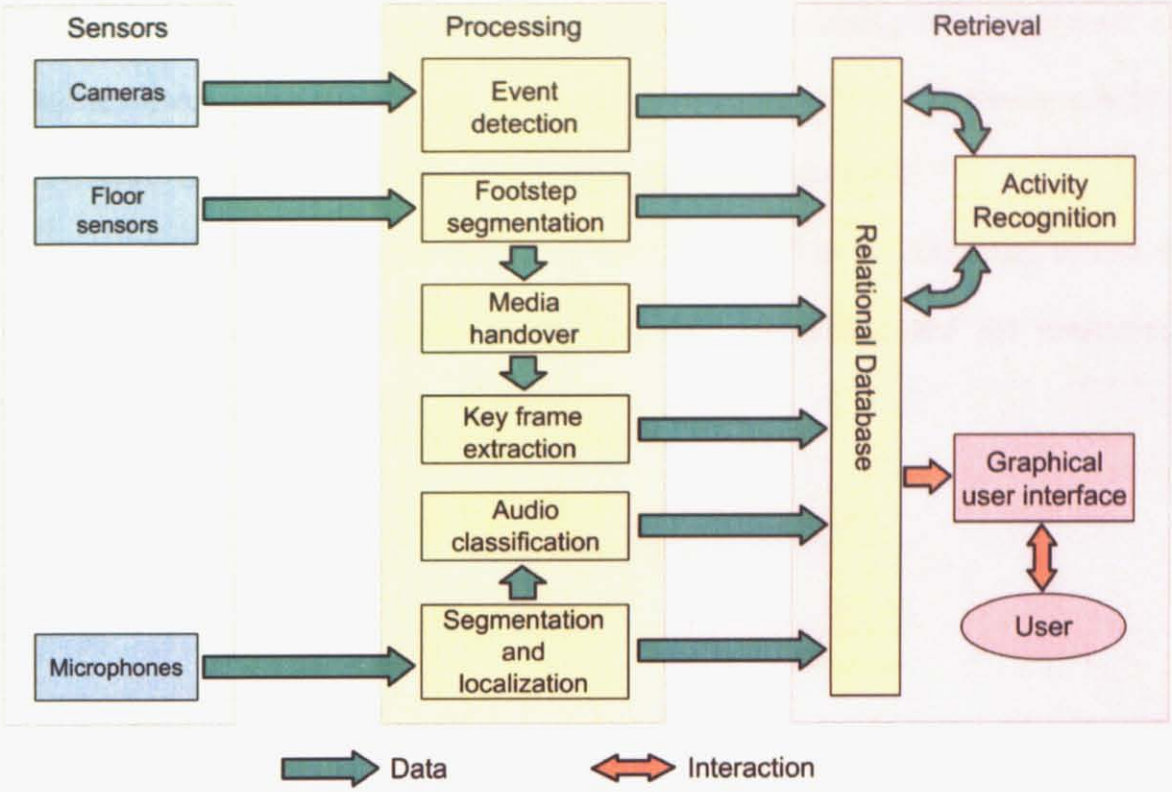

Figure 2. System overview.

residents, retrieve their experiences from the database through a graphical user interface by submitting interactive queries.

4.3 Evaluation

The proposed system consists of a large number of components that function both independently and together to produce results. Therefore, proper evaluation is essential at both component level and system level. Since the system is intended to be used by residents of different age groups in a household, the usability of the system should be high. We evaluate the system using a two-pronged approach. Each functional component is evaluated individually, to ensure that it provides accurate results to the user and the other components using the results. We use standard accuracy measures and experiments where available, while designing experiments and defining new

accuracy measures where necessary. We conduct a user study for the purposes of gathering system requirements and evaluating the overall system. We choose a family who actually lived in ubiquitous home, as the subjects for this study.

The following chapters describe the algorithms used in the functional blocks in Figure 2, the design and implementation of user interaction, and the evaluation experiments conducted.

Chapter 5

Personalized Video Retrieval Using Floor Sensor Data

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.

5.1 Preprocessing

Table 1 shows a subset of the recorded floor sensor data. The entries are ordered according to time. The placing and removal of a foot on the floor will result in one or more pairs of lines. However the pairs may or may not be contiguous, as demonstrated by highlighted rows.

We use a pair-wise clustering algorithm to produce a single data entry, referred to

Table 1: Format of floor sensor data.

Timestamp	X	Y	State
2004-09-03 09:41:20.64	1920	3250	1
2004-09-03 09:41:20.96	2100	3250	1
2004-09-03 09:41:20.96	1920	3250	0
2004-09-03 09:41:21.60	2100	3250	0

Table 2: Format of sensor activation data.

Start time	End time	Duration	X	Y
34880.640	34880.968	0.328	1920	3250
34880.968	34881.609	0.641	2100	3250

as a *sensor activation*, for each pair of lines of input data. Table 2 shows sensor activations corresponding to the data in Table 1. The timestamps are encoded in to a numeric format for ease of programming. The highlighted entry in Table 2 corresponds to the highlighted pair of rows in Table 1.

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.

5.2 Footstep Segmentation

A 3-stage Agglomerative Hierarchical Clustering (AHC) algorithm 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 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

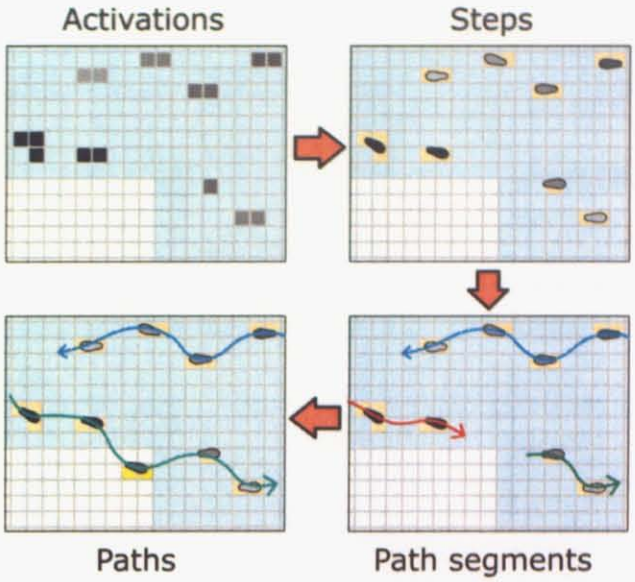

Figure 3: Footstep segmentation.

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.

5.3 Media 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

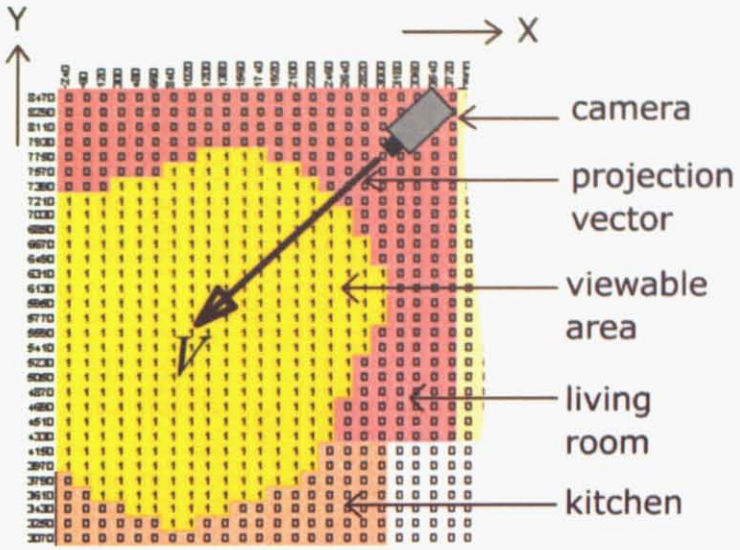

Figure 4: Camera view model.

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*.

In this work we implement two methods for video handover. In the first, we select the camera to view a person based only on his current position. In the second, we try to obtain a frontal view of the person where possible, by calculating the direction of his/her movement.

5.3.1. Camera View Model

To represent the mapping between cameras and their viewable regions, a view model as shown in Figure 4, was constructed for each camera. The projection of the optical axis of the camera on the XY plane, V, is stored as a unit vector. The visibility of a human standing at the location of each floor sensor is represented by the value of 1. This mapping was created by observing images obtained during the experiment. The set of models can be looked up to identify cameras that can see a person at a given position.

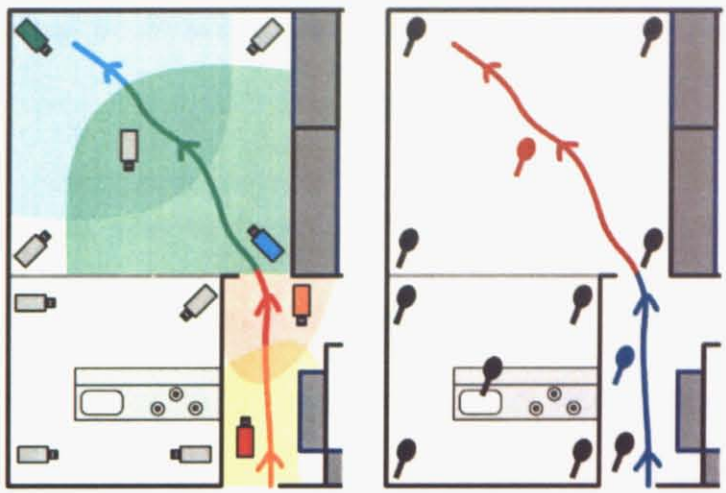

Figure 5: An example of creating a video for a person's path. (a) Position-based video handover, (b) Audio handover.

5.3.2. Position-based Handover

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 5a 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.

5.3.3. Direction-based Handover

This algorithm attempts to select the camera that is most likely to provide a frontal view of the person, when the person is walking inside the house. The direction vector of a

walking person at step p, D_p is estimated by:

$$D_p = \alpha D_{p-1} + (1-\alpha)(X_p - X_{p-1})$$

Here, X_p is the position vector of the step p. The value of α has been empirically set to 0.7 to obtain a relatively smooth direction with steps. The camera to be used is selected by evaluating the scalar product $V.D_p$ for each camera.

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 capture audio from a larger region compared to that seen by 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 5b shows how the microphones are selected for the video clip created in the case of Fig. 5a.

5.4 Key Frame Extraction

We intend to extract a set of *key frames* representing the major content of each video sequence created by media handover. Extracted key frames can provide a compact representation of the video sequence, and can be used for indexing and browsing the sequence in an efficient manner. 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 3. 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.

Table 3: Algorithms for key frame extraction.

Sampling algorithm	Conditions 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)

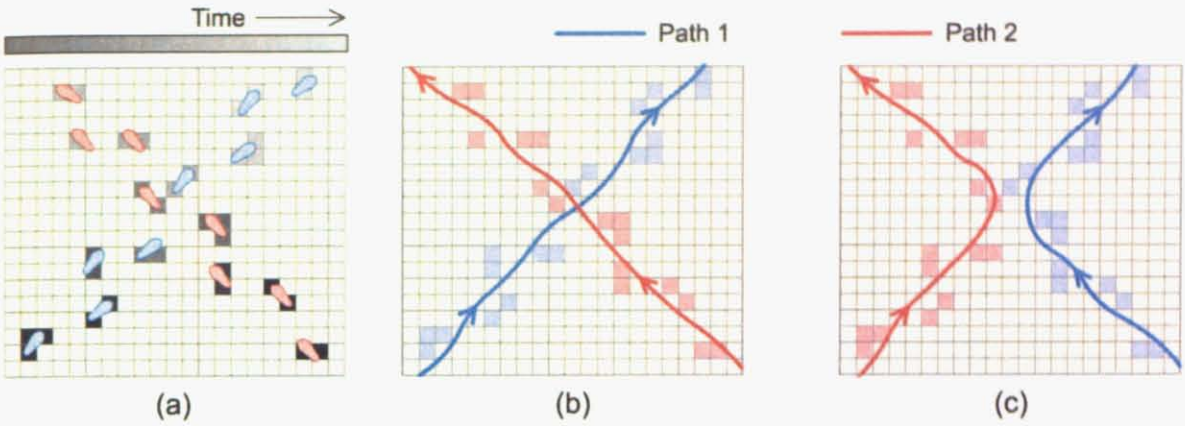

Figure 6: Swapping of paths in footstep segmentation. (a) Sensor data and actual footsteps (b) correct segmentation (c) incorrect segmentation with swapping

5.5 Evaluation of Footstep Segmentation and Media Handover

The hierarchical clustering algorithm for footstep segmentation 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. Figure 6 shows an example of inaccurate segmentation due to swapping paths. The actual footsteps and the sensor activations are shown in Figure 6a. Although the direction of footsteps is considered during clustering to avoid errors, there is a possibility of getting either the correct segmentation (Figure 6b) or an inaccurate pair of paths with swapping (Figure 6c).

The performance of footstep segmentation was evaluated using a data set of approximately 27000 sensor activations, corresponding to 10 hours of data acquisition. Table 3 presents the results of this evaluation. The number of errors present in the results is very small compared to the number of sensor activations and footsteps, despite

the presence of noise, delays, and low resolution. Most of the errors occurred when there were many people in one room and when people entered the areas without floor sensors.

Video clips and key frame sequences that were retrieved using the two methods were evaluated subjectively. Key frame summaries were more effective than video clips when a person stays in the house for a reasonably long duration. Video clips obtained using position-based handover had fewer transitions than those obtained using direction-based handover. For direction-based handover, the calculated gradient is not a robust measure of direction when a person sits and makes foot movements or takes a step back.

Figures 7a and 7b show frames extracted at camera changes for video sequences created using position-based handover and direction-based handover respectively, for the same footstep sequence. The person being tracked is marked by rectangles. It is evident that frame sequences for direction-based handover consist of more key frames, though not necessarily more informative. Position based handover 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 still possible to acquire a frontal view of a walking person most of the time, due to the positioning and orientation of cameras. Therefore, we decided to selected position based handover for camera

Table 4: Results of footstep segmentation.

Description	Value
Number of sensor activations	27020
Total number of paths detected	52
Actual number of paths	39
Number of fragmented paths	15
Number of paths with swapping	4

Figure 7: Subjective evaluation of video handover. (a) position-based (b) direction-based.

selection for personalized video retrieval.

It was possible to create sound tracks with a reasonably uniform amplitude level, using the proposed approach for audio handover. Transitions were smooth, other than for occasional instances where a person was moving from one room to another while talking.

5.6 Evaluation of Key Frame Extraction

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 a novel 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.

5.6.1 Key Frame Extraction Task

The key frame extraction task is based on a video sequence created by position based 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.

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 5.

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 had to answer the following questions:

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

(b) In what ways can it be improved?

5.6.2 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. Additional clarifications were available throughout the experiment, if the subjects needed any. 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.

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

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

Table 5: 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

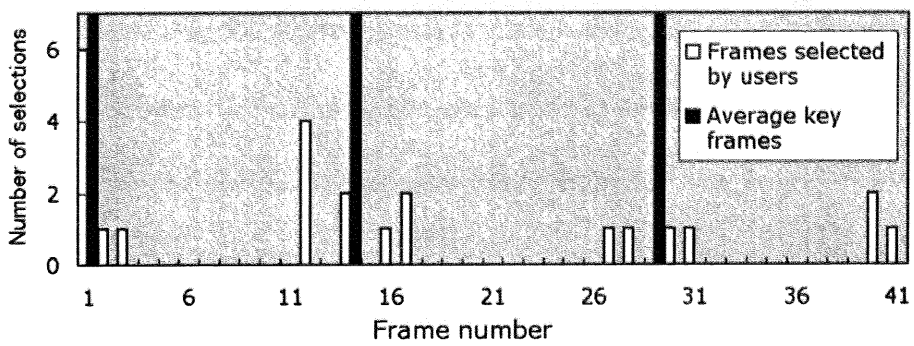

Figure 8: Average key frames.

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:

$$k = \left\lfloor \frac{\sum_{n \in W} nf(n)}{\sum_{n \in W} n} \right\rfloor$$

The average key frames for the frames corresponding to Figure 8 are indicated by black markers on the same graph.

Table 6 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

Table 6: 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

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.

5.6.4 Evaluation of frame sets

The names of the techniques for creating frame sets are abbreviated as shown in Table 7, for ease of presentation.

Table 7: 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

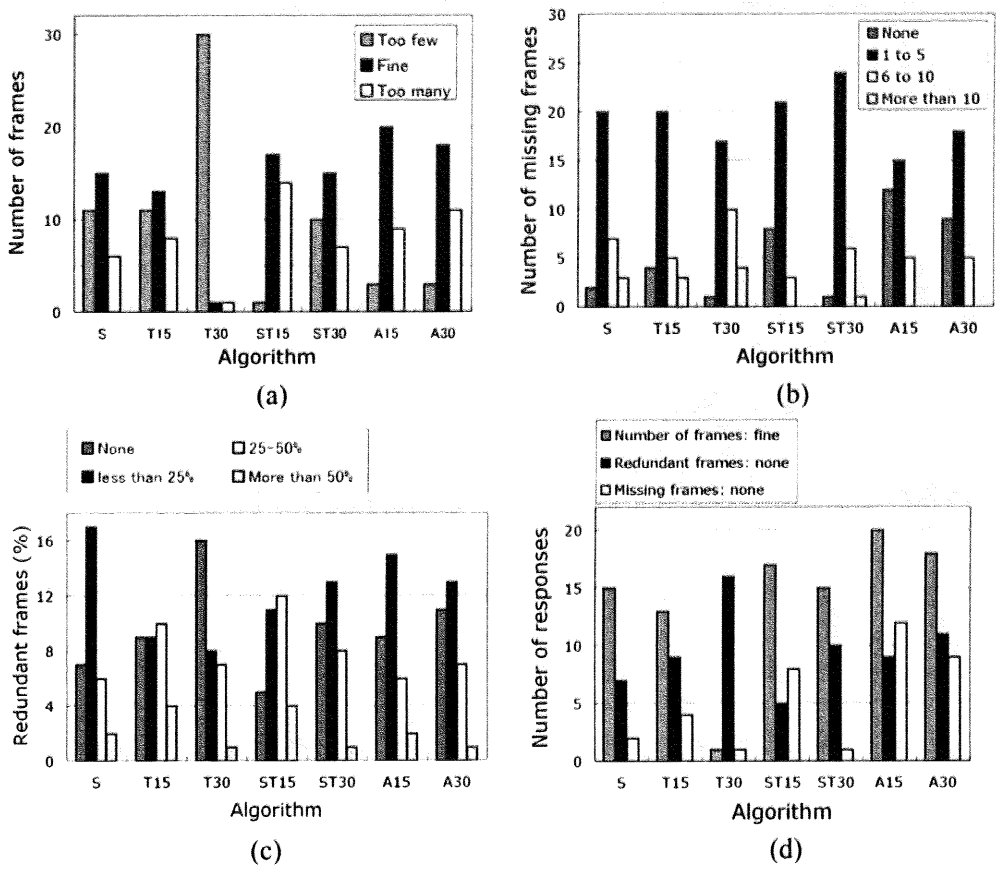

Figure 9: Comparison of votes for the responses. (a) Total number of frames (b) Number of missing frames (c) number of redundant frames (d) Overall comparison.

Figures 9a, 9b and 9c compares the responses from the test subjects for each criterion stated in Table 5. The abbreviations used to denote the algorithms are explained in Table 7. The responses for T30 in Figure 9a 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 9d). It is evident that adaptive sampling has performed

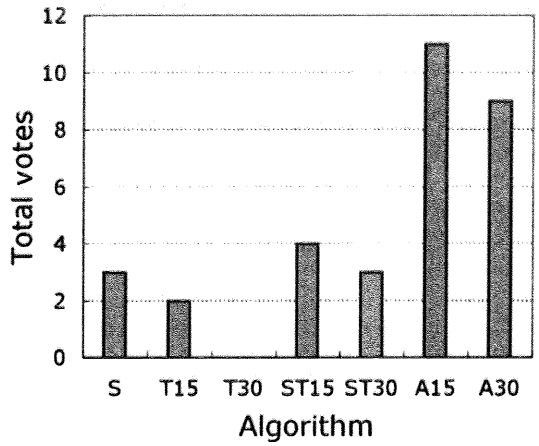

Figure 10: Comparison of votes for the best responses.

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 10 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.

5.6.5 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 11a and 11b show the average key frames and the frame set created by this method respectively, for one sequence. Figure 11c shows the

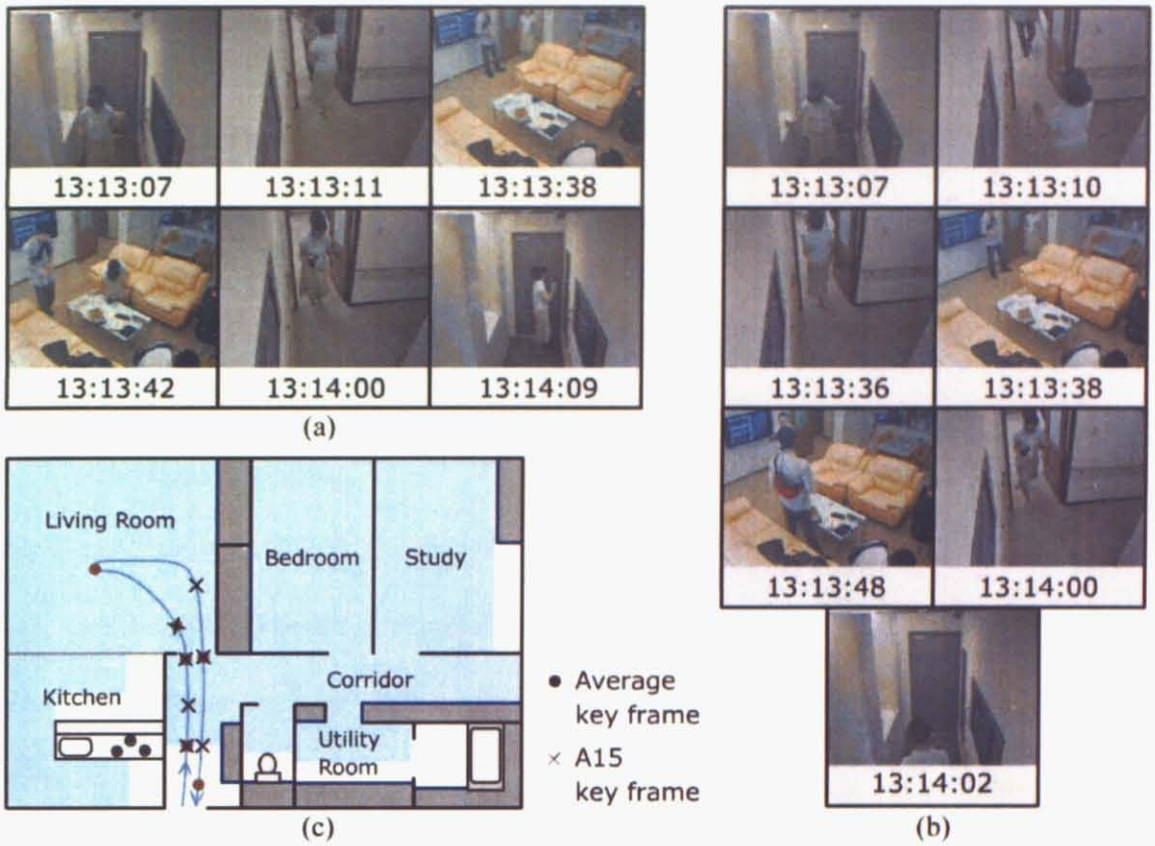

Figure 11: Comparison of average and A15 key frames.

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 12 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.

5.6.6 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)

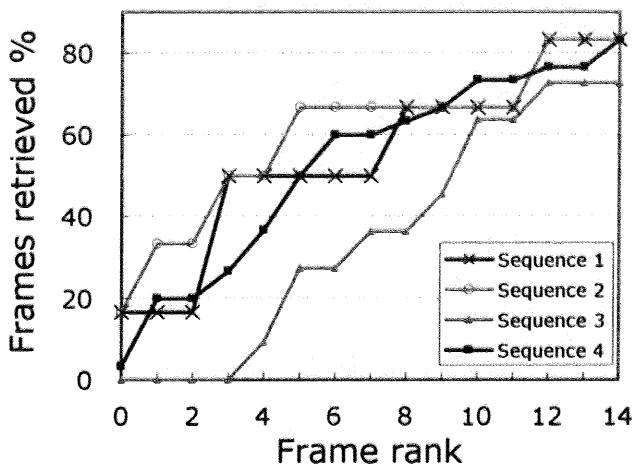

Figure 12: Cumulative performance of key frame extraction.

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.

5.7 Discussion

The floor sensors facilitate tracking people with less computational effort compared to using image analysis. However, they are much more difficult to deploy, compared to cameras. Movement of furniture can generate superfluous data, making tracking difficult. The possibility of using RFID tags together with floor sensors to improve accuracy of tracking is now under investigation.

It is evident that the difference of performance between the two adaptive methods for key frame extraction 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.

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.