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）

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 viii

List of Tables xi

Abstract xiii

List of Publications xvi

Chapter 1: Introduction 1

1.1 Retrieval of experiences from life at home 1

1.2 Motivation 3

1.3 Objectives 3

1.4 Organization of the thesis 4

Chapter 2: State of the Art 5

2.1 Ubiquitous Environments 5

2.2 Multimedia retrieval 7

2.3 Multimedia retrieval in ubiquitous environments 7

2.4 Capture and retrieval of personal experiences 8

2.5 Summary 8

Chapter 3: Ubiquitous Home 10

3.1 Sensors and data acquisition 10

3.2 Data Collection 11

3.3 Issues Related to Capture 12

3.4 Discussion 14

Chapter 4: System Overview 15

4.1 Issues 15

4.2 Outline of the Proposed System 15

4.3 Evaluation 16

Chapter 5: Personalized Video Retrieval Using Floor Sensor Data 18

5.1 Preprocessing 18

5.2 Footstep segmentation 19

5.3 Media handover 20

5.3.1 Camera view model 21

5.3.2 Position-based handover 22

5.3.3 Direction-based handover 22

5.4 Key frame extraction 23

5.5 Evaluation of Footstep Segmentation and Media Handover 25

5.6 Evaluation of Key Frame Extraction 27

5.6.1. Key frame extraction task 28

5.6.2. Experimental procedure 29

5.6.3. Average key frame selection 30

5.6.4. Evaluation of frame sets 32

5.6.5. Comparison with average key frames 34

5.6.6. Descriptive feedback 36

5.7 Discussion 38

Chapter 6: Audio Segmentation for Multimedia Retrieval 40

6.1 Audio Capture and Related Issues 40

6.2 Overview of Audio Analysis 42

6.3 Silence elimination 43

6.4 False Positive removal 44

6.5 Sound source localization 46

6.5.1 Localization based on maximum energy 47

6.5.2 Energy distribution templates 48

6.5.3 Scaled template matching 49

6.6 Audio Classification 51

6.7 Video Retrieval 53

6.8 Evaluation 54

6.8.1. Silence elimination and false positive removal 54

6.8.2. Sound source localization 54

6.8.3. Audio classification 59

Chapter 7: Event and Action Detection Using Multiple Modalities 60

7.1 Issues 60

7.2 Event detection based on lighting changes 62

7.3 Action Classification for Retrieval 65

7.3.1. Clustering of footstep sequences 65

7.3.2. Detailed action classification 66

7.3.3. Combining other modalities to improve accuracy 68

7.4 Evaluation 69

7.4.1. Event detection based on lighting changes 69

7.4.2. Basic activity classification 70

7.4.3. Detailed action classification 71

7.5 Discussion 73

Chapter 8: User Interaction Design 75

8.1 Issues 75

8.2 Approach 75

8.3 Hierarchical Media Segmentation 76

8.4 Interactive retrieval 77

8.5 User interface design 78

8.6 Presentation and Visualization of Results 79

8.6.1. Daily summary 80

8.6.2. Tracked people 82

8.6.3. Key frames 83

8.6.4. Sounds 83

8.6.5. Lighting change events 85

8.6.6. Overall activity visualization 86

8.6.7. Video browser 88

8.7 Example Scenario of Retrieval 88

8.8 Discussion 89

Chapter 9: User Study 91

9.1 Objectives 91

9.2 Participants 92

9.3 Procedure 92

9.4 Results 93

9.5 Discussion 98

Chapter 10: Conclusion and Future Work 100

10.1 Conclusion 100

10.2 Future Work 101

References 103

Appendices 109

A: Material Used For Evaluation of Key Frame Extraction 109

B: Simplified Mathematical Model for Sound Source Localization 125

C: Material Used for the User Study 127

List of figures

Figure 1	Ubiquitous home sensor layout …….	11
Figure 2	System overview ……...	16
Figure 3	Footstep segmentation ……...	20
Figure 4	Camera view model ……...	21
Figure 5	An example of creating a video for a person's path ……..	22
Figure 6	Swapping of paths in footstep segmentation …….	25
Figure 7	Subjective evaluation of video handover ……..	27
Figure 8	Average key frames ……...	31
Figure 9	Comparison of votes for the responses …….	33
Figure 10	Comparison of votes for the best responses ……..	34
Figure 11	Comparison of average and A15 key frames ……	35

Figure 12 Cumulative performance of key frame extraction ……. 37

Figure 13 Microphone positioning and orientation ……... 41

Figure 14 Overview of audio analysis ……... 43

Figure 15 Energy distribution template for the living room …….. 48

Figure 16 Scaled template matching for source localization …….. 50

Figure 17 Microphone positioning and orientation ……... 53

Figure 18 Lighting changes in the living room and the corresponding events .. 62

Figure 19 Threshold estimation using gradient histograms ……... 64

Figure 20 Retrieved events for lighting changes ……... 70

Figure 21 Hierarchical media segmentation …….. 76

Figure 22 Organization of the user interface ……. 79

Figure 23 Visualization of the daily summary …….. 81

Figure 24 Viewing video for tracked people ……. 82

Figure 25 Displaying key frame sets ……. 83

Figure 26 Retrieving video for sound segments …… 84

Figure 27 Interactive retrieval of lighting change events …….. 85

Figure 28 Animated preview of overall activity …… 86

Figure 29 Video browser for multi-camera preview ……. 87

List of Tables

Table 1	Format of floor sensor data ……...	18
Table 2	Format of sensor activation data ……...	18
Table 3	Algorithms for key frame extraction …….	24
Table 4	Results of footstep segmentation ……...	26
Table 5	Criteria for evaluating individual frame sets …….	30
Table 6	Comparison of the number of key frames …….	31
Table 7	Abbreviations for labeling frame sets ……...	32
Table 8	Assignment of microphones to regions …….	45
Table 9	Description of audio database ……...	51
Table 10	Ground truth for audio data ……...	54
Table 11	Overheard sounds before source localization ……	56

Table 12 Overheard sounds after source localization …….. 56

Table 13 Accuracy of sound source localization …….. 58

Table 14 Results of audio classification ……... 58

Table 15 Composition of the activity database …… 67

Table 16 Accuracy of action recognition before using multiple modalities … 71

Table 17 Confusion matrix before using multiple modalities …….. 71

Table 18 Accuracy of action recognition after using multiple modalities …... 73

Table 19 Confusion matrix after using multiple modalities ……. 73

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 Pattern Analysis and Machine Intelligence*.

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.

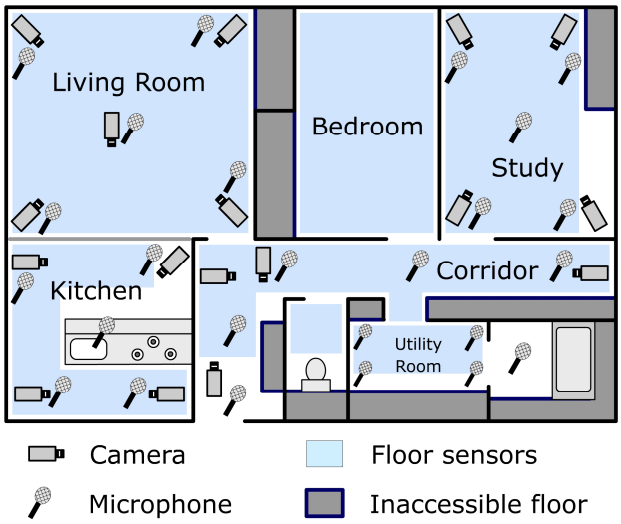


Figure 1: Ubiquitous home sensor layout.

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

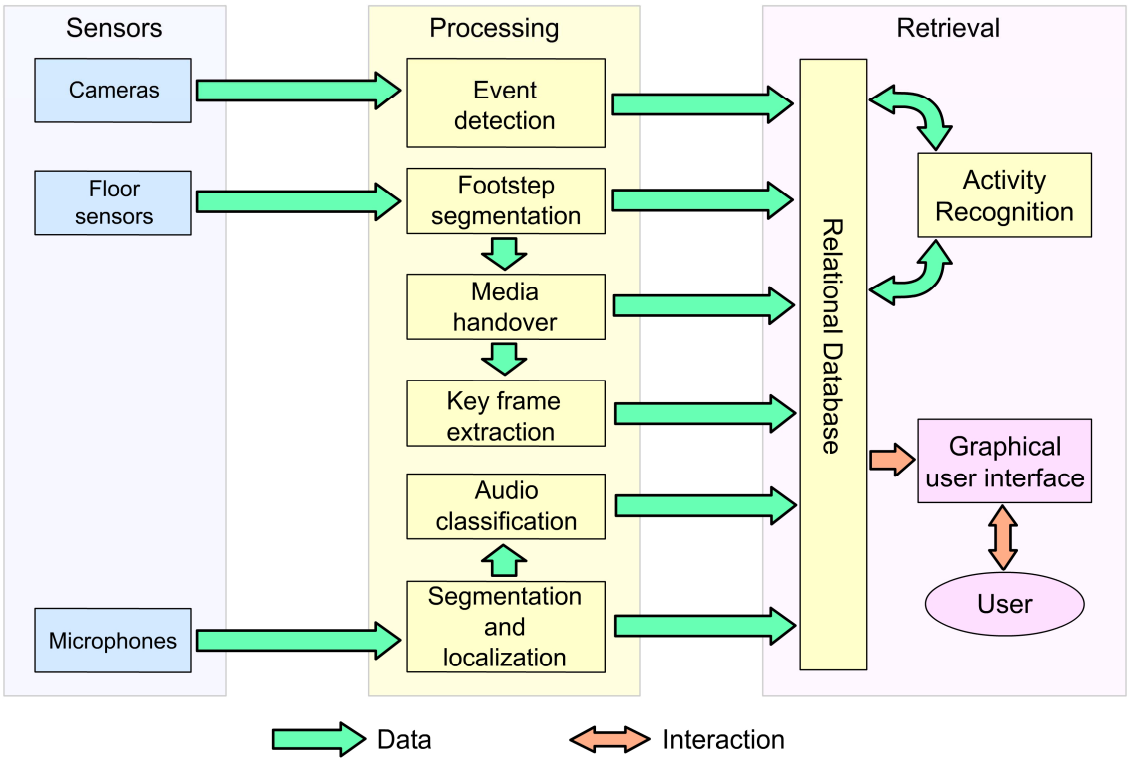


Figure 2. System overview.

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

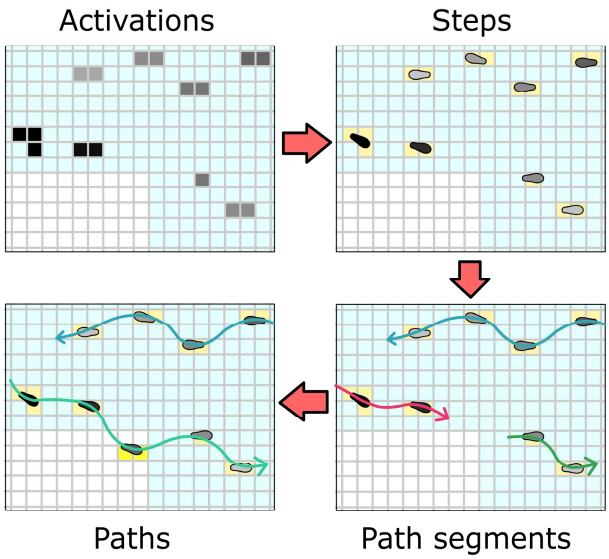


Figure 3: Footstep segmentation.

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

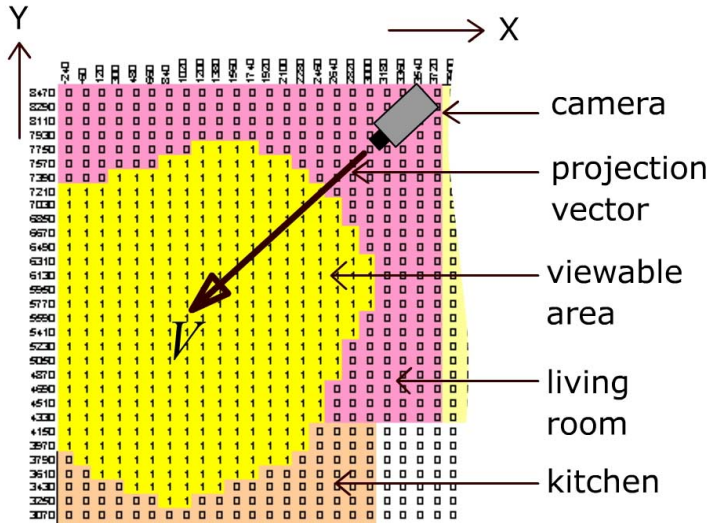


Figure 4: Camera view model.

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.

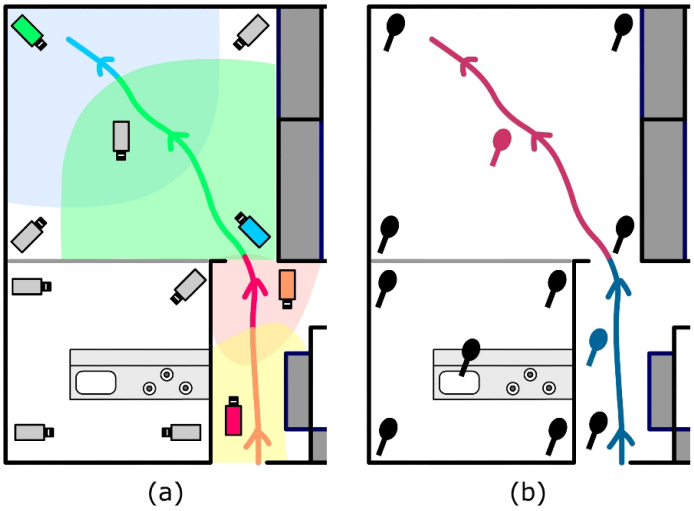


Figure 5: An example of creating a video for a person's path. (a) Position-based video handover, (b) Audio handover.

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:

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

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)

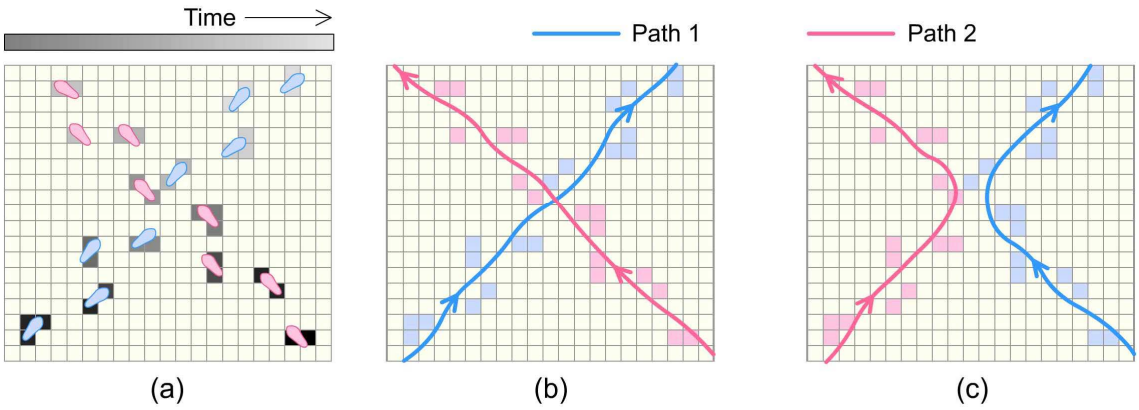


Figure 6: Swapping of paths in footstep segmentation. (a) Sensor data and actual footsteps (b) correct segmentation (c) incorrect segmentation with swapping

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

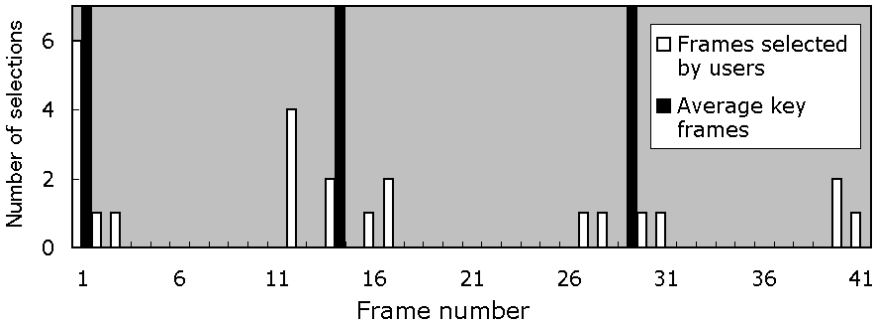


Figure 8: Average key frames.

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:

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

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

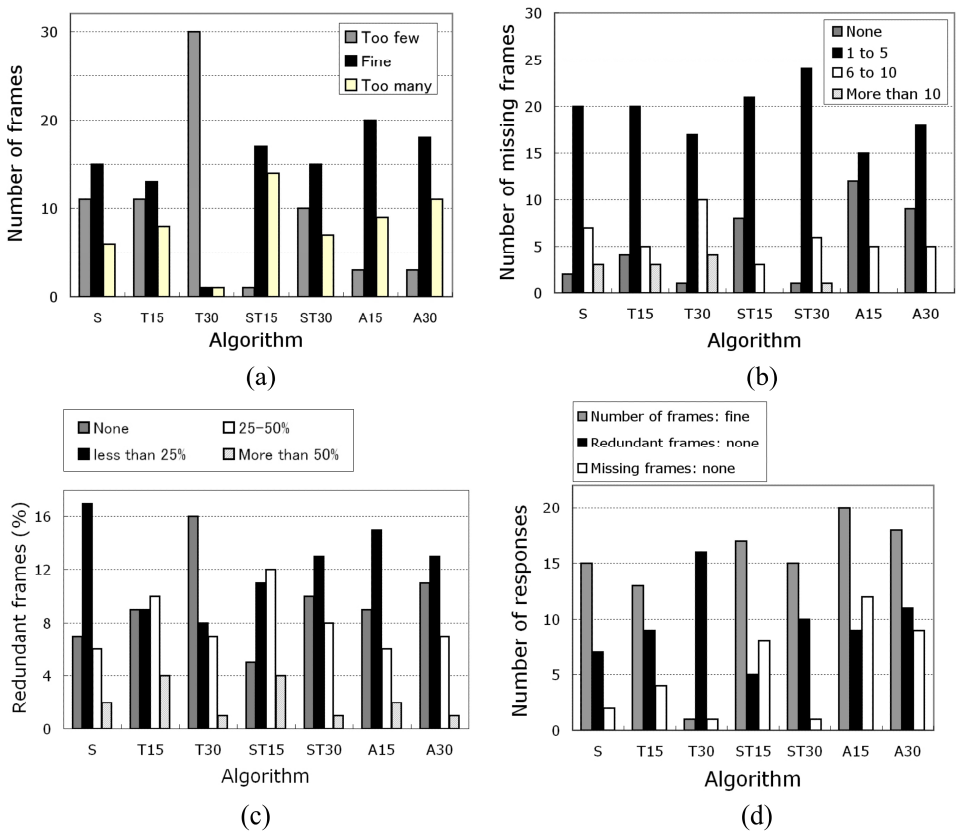


Figure 9: Comparison of votes for the responses. (a) Total number of frames (b) Number of missing frames (c) number of redundant frames (d) Overall comparison.

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

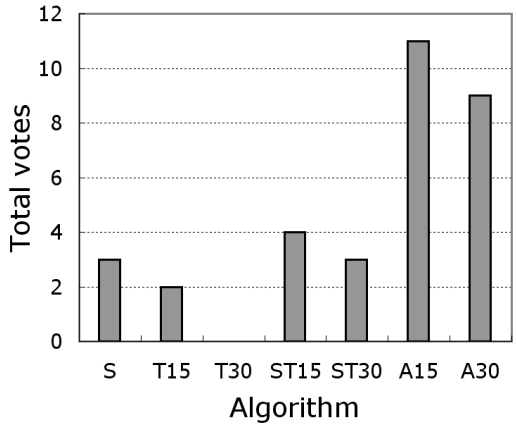


Figure 10: Comparison of votes for the best responses.

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

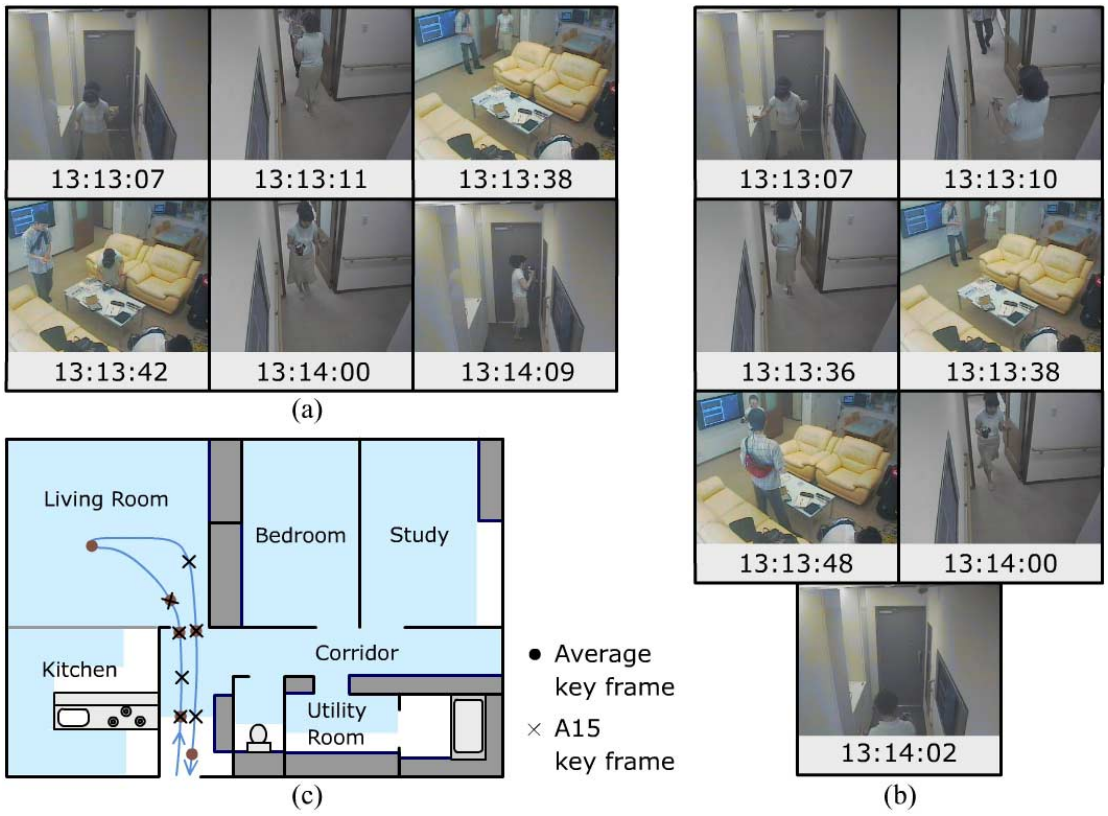


Figure 11: Comparison of average and A15 key frames.

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)

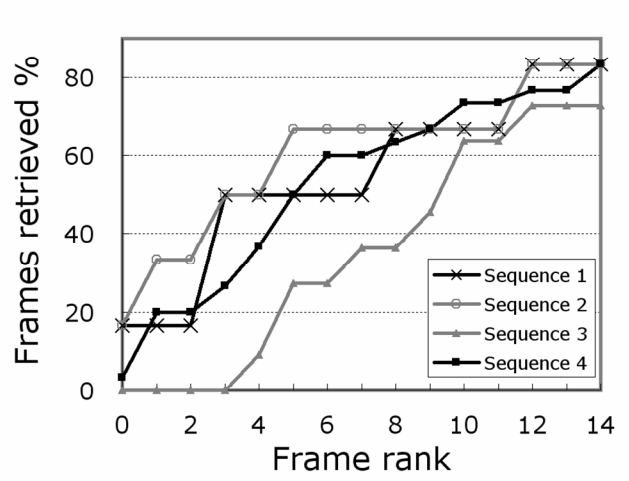


Figure 12: Cumulative performance of key frame extraction.

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.

Chapter 6

Audio Analysis for Multimedia Retrieval

Audio analysis and classification is widely used for video summarization, indexing and retrieval for two main reasons. A large amount of information regarding the content and events of the images and video are contained in the audio signal. Audio data can be successfully used to reduce the search space by selecting cameras according to the results of processing audio data [55]. Being one-dimensional, audio is can be processed with relatively low processing power compared to image data.

The floor sensors are unable to capture data when the people are not treading on a floor area with sensors. Furthermore, they are not activated if the pressure on the sensors is not sufficiently large: for example, when a person is sitting and leaning back with the feet resting on the floor. Audio data can be used to supplement video retrieval in such situations. Audio analysis can also be conducted independently, to retrieve multimedia related to events that are characterized by sound (e.g.: conversations). The remaining sections of this chapter describe our use of audio analysis for multimedia retrieval from ubiquitous home.

6.1 Audio Capture and Related Issues

Cardioid, uni-directional microphones are mounted along the corridor and at the corners of the rooms. An Omni-directional microphone is installed at the center of each room where data are captured. Figure 13a shows the positioning and orientation of the microphones. The numbering of the microphones will be used to refer to them in the

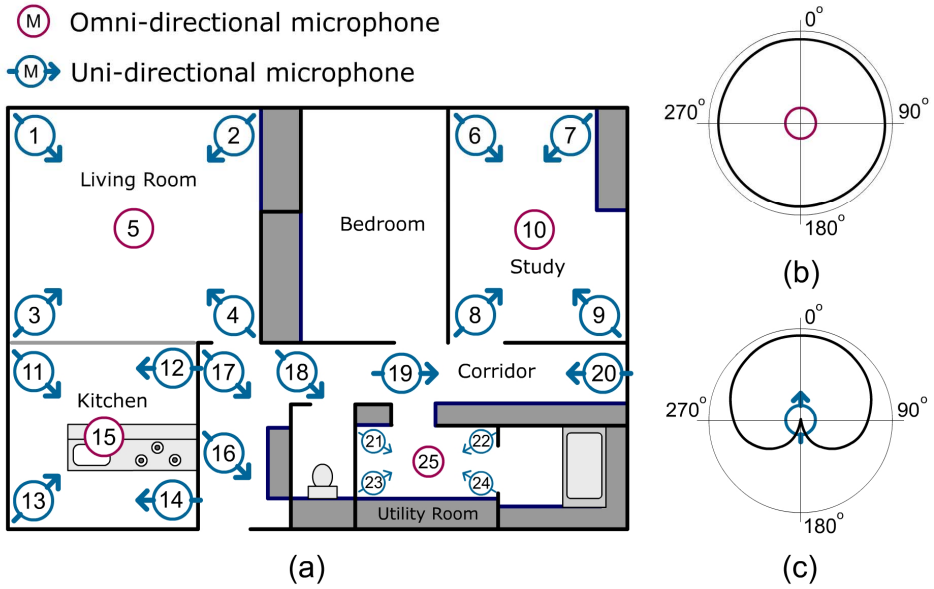


Figure 13: Microphone positioning and orientation.

coming sections of the paper. Figures 13b and 13c show the directional responses of omni directional and cardioid microphones respectively.

With 25 microphones recording continuously, the amount of audio to be processed is fairly large. The coverage of a microphone is much less restricted for a microphone than for a camera. Sounds from one source being picked up by several microphones, even those outside the room they are mounted in. Some form of sound source localization is therefore necessary to select cameras in order to show what caused a particular sound.

A brief review of the techniques used in sound source localization can be found in [56]. The most common approaches are based on time delay of arrival (TDOA), microphone arrays [57] and beam-forming techniques [58]. However, the conditions required by these approaches are not satisfied by the microphone setup in ubiquitous

home. For example, the small size of the rooms and reflections from walls can adversely affect the performance of TDOA based techniques.

There has been some recent research on sound source localization for home-like ubiquitous environments. Vacher et al. use audio from multiple microphones to facilitate tele-monitoring of patients based on audio events by selecting cameras based on maximum audio energy [59]. Bian et al. [60] investigate sound source localization in a home-like setting, using TDOA and microphone quads. However, both these techniques are designed only for situations where only a single sound source is active at a given time.

Since the microphones are located in close proximity, redundancy in captured audio data is fairly high. A trade-off has to be made between utilizing the redundancy to improve the accuracy of retrieval and minimizing processing by removing redundancy.

6.2 Overview of Audio Analysis

Figure 14 is an outline of the system we propose for video retrieval based on audio analysis. The audio streams are synchronized and partitioned into *segments* of one second each. Sets of audio segments that are captured using different microphones during the same 1 s interval (hereafter referred to as *segment sets*) are processed together. We start by eliminating silence from audio signals captured by individual microphones. The resulting sound segments are processed further to reduce false segments due to noise. The next step is to identify the location/s of the sound sources for each sound segment. The results are used to retrieve video directly and also classified into different classes of sound, which can again be used for video retrieval. The following sub-sections describe these steps in detail.

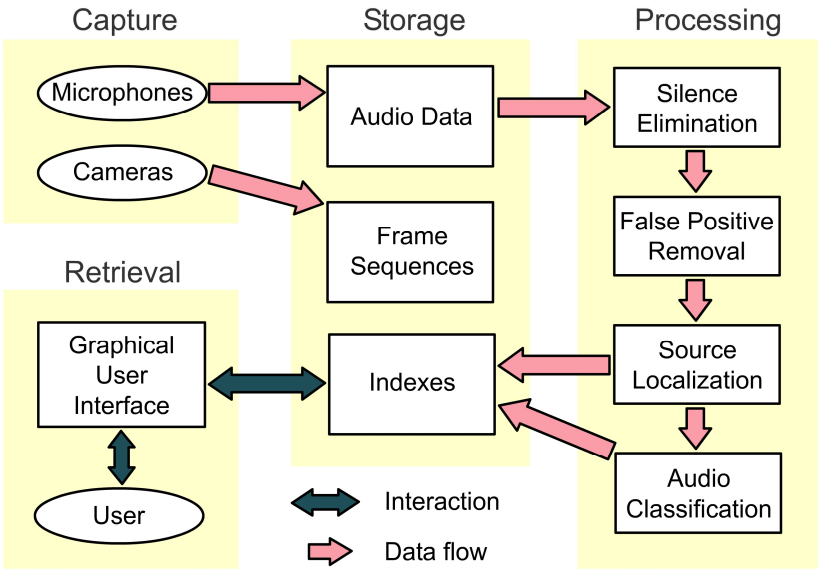


Figure 14: Overview of audio analysis.

6.3 Silence Elimination

A common approach for silence elimination for a single audio stream is to compare the RMS power of the audio signal against a threshold value [29]. We select this approach since it is simple, and adequate for stationary microphones in a controlled environment such as ubiquitous home. The threshold for each microphone is estimated by analyzing audio data for silence and noise for that microphone. Audio segments with a total duration of one hour were extracted from different times of day. These were partitioned into *frames* having 300 samples. Adjacent frames had a 50% overlap. The Root Mean Square (RMS) value of the samples in each frame is calculated and recorded, and the statistics obtained for each clip.

Since the probability distribution of the RMS values for different audio clips were not significantly different, the data were combined to make a single probabilistic model for silence and noise. The threshold value was selected to be at 99% level of confidence

according to this distribution. The value was selected below 100% as false negatives (sound misclassified as silence) are more costly than false positives (silence misclassified as sound). The latter can be eliminated using further analysis.

For silence elimination using the above threshold value, each audio segment is divided into overlapping frames in the same manner as previously, and the RMS value of each frame is calculated. If the calculated RMS value is larger than the threshold, the frame is considered to contain sound. Otherwise, it is considered to contain silence. Sets of contiguous frames with duration less than 0.1s are removed. Sets of contiguous frames that are less than 0.5s apart are combined together to form single segments.

The first stage of silence elimination is based on individual microphones. The audio stream is divided into overlapping frames in the same manner as previously, and the RMS value of each frame is calculated. If the calculated RMS value is larger than the threshold, the frame is considered to contain sound. Otherwise, it is considered to contain silence. The total duration of contiguous frames with RMS above the threshold are used for further processing. Sets of contiguous frames with duration less than 0.1s are removed. Sets of contiguous frames that are less than 0.5s apart are combined together to form single segments.

6.4 False Positive Removal

The second stage uses the data from multiple microphones in close proximity to reduce false positives resulting due to noise. For this purpose and for use in the following stages, the microphones are grouped in to regions as specified in Table 8. The bedroom has no microphones installed. However, it was identified as a separate region, for use in further analysis.

For each microphone, a binary sound segment function $B(n)$ and cumulative sound segment function $C(n)$ are defined by

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

$B(n) = 0$ otherwise

$C(n) = \sum B(n)$ for the set of microphones in the same room.

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

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

$S(t) = 0$ otherwise

where

$\otimes$ denotes convolution,

Table 8: Assignment of microphones to regions.

Region	Label	Microphones
Living room	LR	1-5
Study room	SR	6-10
Kitchen	KT	11-15
Entrance	EN	16,17
Corridor	CR	18-20
Utility room	UR	21-25
Bedroom	BR	-

$M(t) = [1 \quad 1 \quad 1]$ and

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

The value $S(t) = 1$ indicates that a sound was heard in the region during the t^{th} second, whereas $S(t) = 0$ indicates that no sound was heard within the region during the t^{th} second.

The set of sound segment functions are passed as input to the next stage, together with the audio data. Only the sound segments where $S(t) = 1$ will be processed in the following stages.

6.5 Sound Source Localization

After noise elimination, we have a set of sound segments for each microphone in a given region, for the situations where there was a sound *heard* in that region. We categorize the sounds contained in these segments into two types. One is *local* sounds, that is, sounds generated in the same region as the microphone belongs to. The other, *overheard* sounds, refers to the sounds that are generated in a region other than that the microphone belongs to. Each segment can contain either, or both of these types.

We intend to remove sound segments that contain only overheard sounds, from the results of noise removal. Such segments, if not removed, can mislead algorithms for video retrieval. We refer to this task as *sound source localization*, as it identifies the regions where one or more sound sources are present. However, it should be noted that we do not wish to identify at this stage whether there is a single source or multiple sources in each region. Furthermore, our source localization is restricted to region level, not to the exact location.

The regions of the house are partitioned in different ways. Some places, such as the study room, are separated from the rest by a door, and only a few sounds propagate to the other regions and rooms. However, the situation is different for some other regions such as the living room and kitchen, which are not so strongly partitioned. The situation is complicated further due to sounds from the bedroom being heard outside, while the bedroom itself does not have any microphones installed.

6.5.1. Localization Based on Maximum Energy

A simple approach for sound source localization is to select the microphone that captures the sound with the highest volume and selecting the region associated with it [56]. Although this will miss out some sound sources when there are multiple sound sources emanating sound in the same 1 s interval, we investigate the performance of this approach for the purpose of comparison. Based on this approach, we design the following algorithm for *localization based on maximum energy*.

1. For the current segment set, calculate the mean square value (which is proportional to short-term energy) of the samples in each segment.
2. Calculate the average energy for each region by averaging mean square values for the segments from the microphones in that region.
3. Select the region that has the maximum value, as the region where the sound source is located.

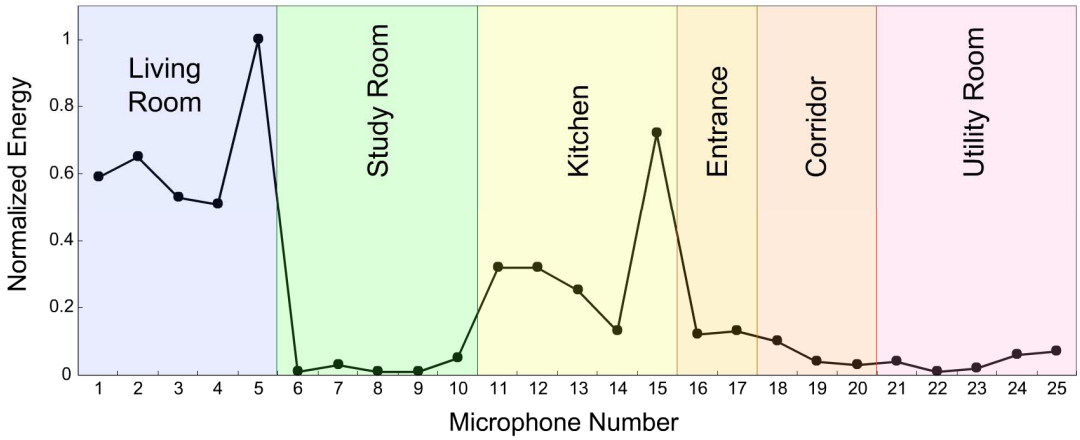


Figure 15: Energy distribution template for the living room.

6.5.2. Energy Distribution Templates

A sound generated in one region of ubiquitous home can be heard in other regions, with varying levels of intensity. Based on this fact, we attempt to model local sounds in each region with the variation of energy received by each microphone. For each region r, a number of segment sets are selected for instances when sounds were generated only in that particular region. The energy distribution $E(n)$ for the segment set is determined by calculating the energy for each segment in the set. The *Energy Distribution Template*, T_r of the region r is estimated by averaging all the energy distributions. Each template is normalized to be in the range [0 1]. Figure 15 shows the template for sounds originating in the living room. Normalized energy is quite high for the microphones in the living room compared to those in other regions, with the highest energy level recorded by the omni directional microphone. However, relatively high levels of energy are registered by the microphones in the kitchen, due to the absence of a wall between the kitchen and the living room.

6.5.3. Scaled Template Matching

Each audio segment set as a mixture of audio signals generated in one or more regions of the house. We hypothesize that the energy distribution of a segment set is a linear combination of one or more energy distribution templates, and attempt to identify them (see *Appendix B* for details). We use the following *scaled template matching algorithm* for finding these templates. The main idea behind the algorithm is to repeatedly identify the loudest sound source available, and removing its contribution. This process repeats until it is evident that there is no significant sound energy left to assume the presence of a sound source. The procedure for scaled template matching is described below:

1. Calculate the energy distribution, $E(n)$ for the current segment set
2. For each region r, determine average energy E_r by averaging energy values of the microphones in that region
3. Find the region r in the distribution with the maximum value of E_r for the current energy distribution $E(n)$. Identify this region as containing a local sound segment.
4. Scale $E(n)$ by dividing by the max value in that region, A_m.
5. Subtract the template T_r corresponding to this region, r, and obtain the residual $R(n)$
6. If the average value of $R(n) \leq 0.2$ then stop. Otherwise, multiply $R(n)$ by A_m again
7. Repeat steps 2-5 on $A_m R(n)$, for k times where k is the number of regions where sound segments are detected.

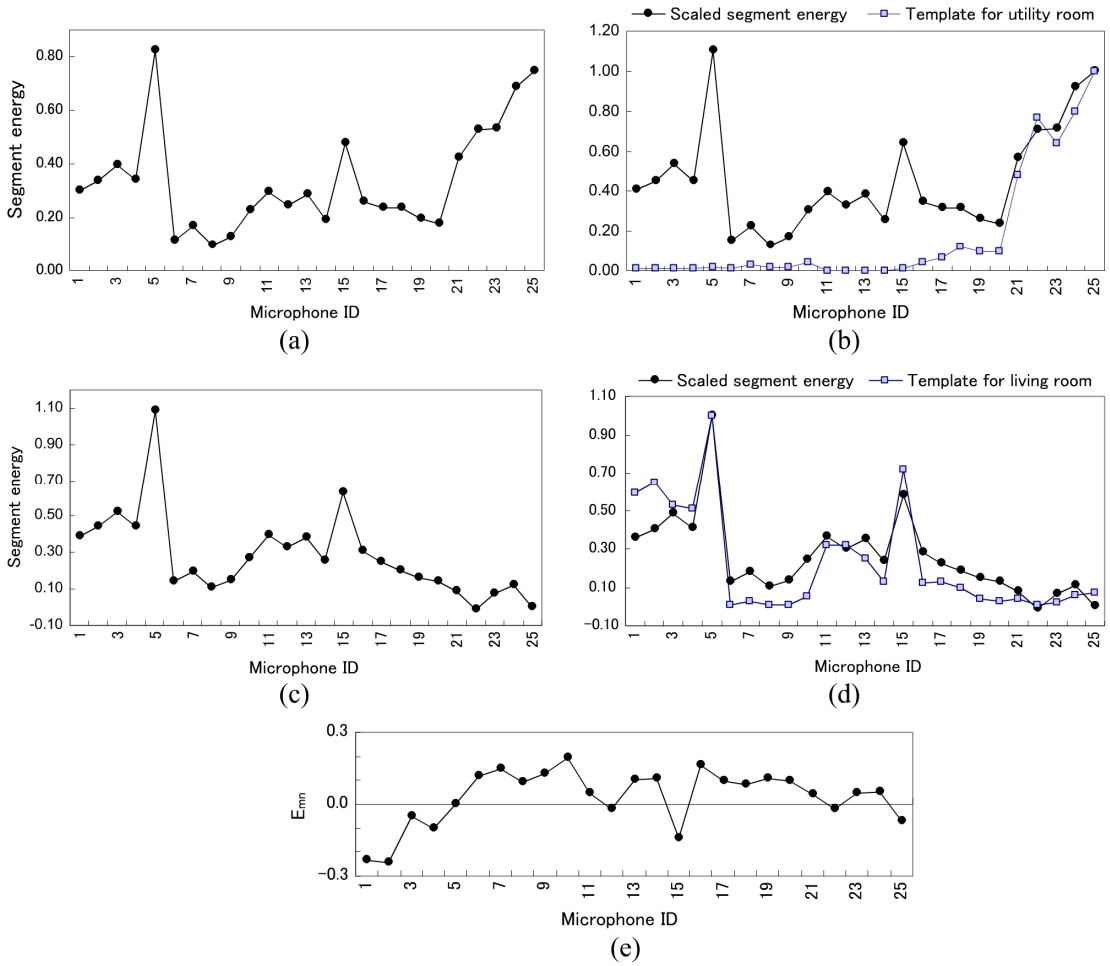


Figure 16: Scaled template matching for source localization. (a) original segment set (b) matching with template for the utility room (c) energy distribution after template subtraction (d) matching with template for the living room (e) residual segments

Figure 16 visualizes the application of this algorithm to a set of audio segments. In this case, there have been local sound segments in the utility room and the living room. Figure 16a shows the energy distribution for the current segment set. Since the highest average energy is present in the *utility room* region (microphones 21-25), the utility room is identified as having local sounds. The energy distribution is scaled by diving by the energy for microphone 25 (the maximum value for this region), and the template for utility room is subtracted from the result (Figure 16b). The result after subtraction, $R(n)$, is shown in Figure 16c. The process is repeated for the template for

the living room (Figure 16d), which records the highest average in $R(n)$. The algorithm stops at this point as the average value becomes lower than the set threshold (Figure 16e).

6.6 Audio Classification

The results of source localization can be used to retrieve video for basic *audio events*, that is, instances where something generated a sound in a given region. However, this retrieval is at a very low semantic level and will results in a large amount of video. Retrieval by different classes of audio, such as voices and music, will greatly enhance the precision of retrieval. We conducted a pilot study on audio classification. An audio database was constructed by studying the sound segments extracted from experiments in ubiquitous home. These were classified into the categories shown in Table 9. The duration of each audio clip in the database was between 1 and 15 seconds.

The classes were selected by observing data from the real-life experiment, and

Table 9: Description of audio database.

Label	Class	No. of audio clips
1	Footsteps	40
2	Noise	40
3	Voices of people inside the house	112
4	Voice of a household robot	32
5	Voices from television	50
6	Other sounds from television	60
7	Vacuum cleaner	60
8	environmental sounds	86
	Total	480

aiming at detecting higher level events such as conversations, and watching TV. We attempted audio classification based on frame-based and clip-based time domain features. These features are relatively easy to calculate given the large amount of data, and facilitate reasonably accurate classification according to results reported in similar work [61]. The frame size for calculation of features was the same as for silence elimination. The selected features were:

- Mean of RMS values of the set of frames
- Standard deviation of RMS values of the set of frames
- Mean of Zero crossing ratios of the set of frames
- Standard deviation of Zero crossing ratios of the set of frames
- Silence ratio of the audio clip

All of the features were calculated according to the definitions in [61]. Each feature was normalized by subtracting the mean and dividing by the standard deviation for the feature in the entire database.

A number of classifiers, including *Multi-layer Perceptron (MLP)*, *k-Nearest Neighbor* and *Random Forest* were trained and tested on the database. The results were evaluated using 10-fold cross validation. *AdaBoost with MLP* classifier yielded the highest overall accuracy, and therefore was selected as the classifier to be implemented in the proposed system.

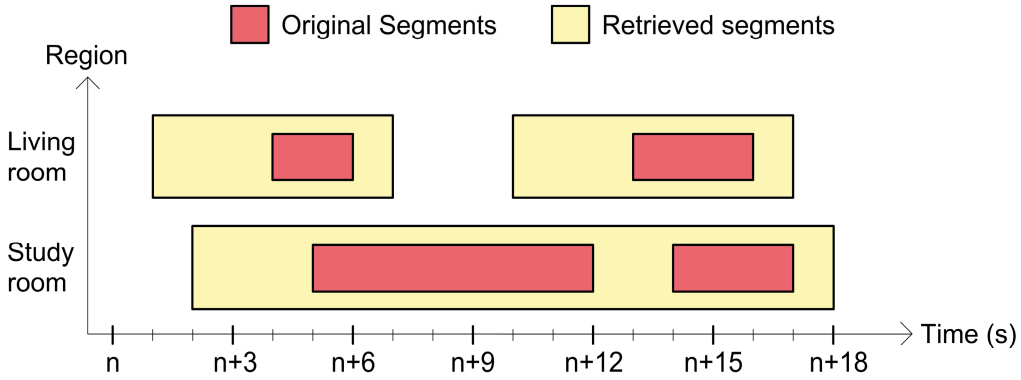


Figure 17. Microphone positioning and orientation.

6.7 Video Retrieval

The method of audio-based retrieval is fairly straightforward once audio segments are localized. First, consecutive sound segments that are less than 4 seconds apart are joined together, to prevent retrieving a large number of fragmented videos for the same event. After joining, the start time for retrieving video is set to be 3 seconds earlier than that of the sound segment, to allow the user to prepare to receive the audio event. The end of the video clip is set to be 1 second later than that of the sound segment. Figure 17 is a visual representation of this process. Video clips are retrieved from all cameras in the region for the time interval determined as above. The same approach is applicable for retrieving video segments for different classes of audio.

The video clips retrieved for footstep sequences are extended using sound segments as follows. If there is only one footstep sequence overlapping partially with a sound segment, it is combined with the footstep sequence. The video created by video handover is extended to include the time during which sounds were present before the start of the footstep sequence, or after the sequence ended. This improves the video in

certain situations, such as when a person enters a region without floor sensors but continues to talk.

6.8 Evaluation

6.8.1 Silence Elimination and False Positive Removal

The performance of silence elimination and false positive removal was evaluated using 90 hours of audio data, extracted from the data captured during a single day in ubiquitous home. Silence elimination resulted in 0% false negatives and 2.2% false positives on this data set. The algorithm for false positive removal was able to remove 83% of the false positives that remained after silence elimination in individual audio streams.

6.8.2 Sound Source Localization

We used 200 minutes of audio data captured during the real-life experiment, for evaluation of sound source localization. The data were captured from 7:45 a.m. to 11:05 a.m. on the 12th April 2005, from all microphones. This time interval was selected to ensure that all regions of the house were used for a considerable duration for ordinary household activities. The data were transcribed manually to find out the ground truth with regard to the sounds generated in each region (Table 10).

We studied the degree of overhearing between different regions of ubiquitous

Table 10: Ground truth for audio data.

Region	LR	SR	KT	EN	CR	UR	BR
No. of segments	5238	639	1667	134	192	814	> 105

home, after applying silence elimination and noise removal on these audio data. To present the situation with overheard sounds in a meaningful way, we calculate the ratio of overhearing, H_{ij} as

$$H_{ij} = N_{ij} / T_j$$

Where

N_{ij} = Sound coming from room j and heard in room i

T_j = Total number of sound segments generated in room j

and $i, j \in \{LR, SR, KT, EN, CR, UR, BR\}$

Table 11 shows the 7×7 matrix $H = [H_{ij}]$. To take an example on how to interpret the values, 73% (0.73) of the sounds coming from the living room can be heard in the kitchen. It is evident that the results are not exactly symmetric. For example, only 28% of the sounds coming from the kitchen can be heard in the living room, compared to the 73%for the other way. The reason is that most sounds generated in the living room are loud sounds, for example group conversations and sounds from the TV. The sounds generated in the kitchen, on the other hand, are softer (e.g. preparing food).

For maximum energy-based localization, it is not logical to calculate matrix H as the algorithm can detect only one region with local sounds for a given segment set. Table 12 shows the situation with overheard sounds after source localization using

Table 11: Overheard sounds before source localization.

Region	LR	SR	KT	EN	CR	UR	BR
LR	1.00	0.00	0.28	0.03	0.13	0.00	0.00
SR	0.00	1.00	0.00	0.01	0.39	0.07	> 0
KT	0.73	0.00	1.00	0.12	0.01	0.00	0.00
EN	0.06	0.00	0.10	1.00	0.36	0.16	> 0
CR	0.00	0.02	0.02	0.27	1.00	0.18	> 0
UR	0.00	0.00	0.00	0.02	0.27	1.00	0.00
BR	?	?	?	?	?	?	?

Table 12: Overheard sounds after source localization.

Region	LR	SR	KT	EN	CR	UR	BR
LR	1.00	0.00	0.03	0.02	0.03	0.00	0.00
SR	0.00	1.00	0.00	0.00	0.04	0.00	0.00
KT	0.05	0.00	1.00	0.00	0.00	0.00	0.00
EN	0.01	0.00	0.00	1.00	0.06	0.05	0.00
CR	0.00	0.04	0.00	0.13	1.00	0.05	> 0
UR	0.00	0.00	0.00	0.02	0.06	1.00	0.00
BR	?	?	?	?	?	?	?

scaled template matching.

We define the Precision P, Recall R, and Balanced F-measure F for audio segmentation for each region of the house, as:

$$P = N_c / N_t$$

$$R = N_c / N_a$$

$$F = 2PR/(P+R)$$

where

$N_c =$ no. of local audio clips retrieved correctly

$N_t =$ total no. of clips retrieved

$N_a =$ actual no. of local audio clips

Table 13 summarizes the results of the evaluation. The precision, recall and balanced F-measure are shown for original data after false positive removal, for the results using maximum energy based selection and for the results obtained by scaled template matching. The values cannot be determined for the bedroom since ground truth is not known due to the absence of microphones.

It is evident that the results for rooms other than the kitchen have improved to near 100% with the scaled template matching algorithm. For the kitchen, there has been a significant improvement even though the results are not as good. The high recall rates demonstrate the ability of the algorithm to localize multiple sound sources with a high accuracy.

The high accuracy recorded with the proposed scaled template matching algorithm is mainly due to the fact that it utilizes the high degree of partitioning present

Table 13: Accuracy of sound source localization.

	Before localization			Max. energy			Proposed method		
Region	P	R	F	P	R	F	P	R	F
LR	0.56	1.00	0.72	0.97	0.93	0.95	0.95	0.99	0.97
SR	0.78	1.00	0.88	1.00	0.45	0.62	0.96	0.80	0.88
KT	0.72	1.00	0.84	0.96	0.78	0.86	0.97	0.79	0.87
EN	0.69	1.00	0.82	1.00	0.60	0.75	0.86	0.89	0.88
CR	0.46	1.00	0.63	1.00	0.80	0.89	0.84	0.97	0.90
UR	0.71	1.00	0.83	1.00	0.82	0.90	0.91	0.89	0.90

Table 14: Results of audio classification.

Class	1	2	3	4	5	6	7	8
Precision	0.68	0.91	0.79	0.91	0.73	0.82	1.00	0.78
Recall	0.80	1.00	0.82	0.97	0.76	0.88	1.00	0.67
F-measure	0.74	0.95	0.81	0.94	0.74	0.85	1.00	0.72

in a home-like environment. Different results can be expected in environments that are larger and have less partitioning. The evaluation of 200 minutes of audio from each of the 25 microphones was quite tedious. While this was done with care to ensure as less error as possible, there may still be some human error (of about 1-2%) included in the results.

6.8.3 Audio Classification

The results of audio classification are presented in Table 14. The accuracy of classification using only time domain features suggest that more accuracy can be obtained by adding frequency domain or MFCC domain features. The classes in the sound database have to be refined further, after a study of requirements for retrieval based on audio classification.

Chapter 7

Event and Action Detection Using Multiple Modalities

The algorithms described in the previous chapters facilitate efficient multimedia retrieval by automated selection of sources, video summarization and audio classification. These results can be interpreted as corresponding to basic events. The video clips and key frames retrieved using footstep sequences correspond to "human presences". However, this event is quite coarse in terms of granularity, as the person can be walking, standing, or performing several other tasks during his/her presence. The video retrieved for different audio classes correspond to "audio events" of the same class, in the corresponding location; for example, "sounds from the television in the living room", "vacuum cleaning the study room". As for the second example, it can also be interpreted as an action.

The rest of this chapter describes the techniques we use to detect additional events using image data (which were hitherto unprocessed) and to segment the video clips retrieved for footstep sequences based on the activities, or *actions*, performed by the person.

7.1 Issues

Both action recognition and event recognition are hard, owing to two main problems. The first problem is related to definition of an action or an event. There has been an ongoing discussion on the question "what are actions and events in the context of multimedia" over the last decade [62][63]. Some actions can be interpreted as events,

and vice versa, further complicating the issue. Since it is hard to define what events and actions are, identifying a complete set of events or activities is a difficult task. The usual approach to solve this problem is to recognize a desired set of actions or events against "others" [64]. An alternative, which is particularly useful in surveillance, is to detect "unusual actions and events" [65].

The second problem is the recognition of desired actions and events with the available sensor data. The common approach used to solve this problem is to build a classifier based on supervised learning using a set of training data. However, most of the time, the sensor data are not sufficient to represent the desired action or event, unless certain preconditions are satisfied. For example, most of the researches on sports video retrieval rely on the common camera positioning for video capture, and context data representing the play field [66]. Furthermore, the features extracted from the sensor data might not contain information that facilitates easy and accurate classification, unless the features are selected carefully. Therefore, accurate recognition of actions and events with high semantic levels is a difficult task.

We use a *data-driven* approach in action and event detection for the ubiquitous home. The actions and events are selected by observing the size and dimensionality of data, and also the content. We choose simple actions and events that are easy to detect, yet helpful to the user to interpret the results. While the algorithms described in the previous sections were based on analyzing a single sensor modality, we combine data from multiple sensory modalities where they compliment each other, for more accurate action recognition. The following sections describe the algorithms used for action and event detection in ubiquitous home.

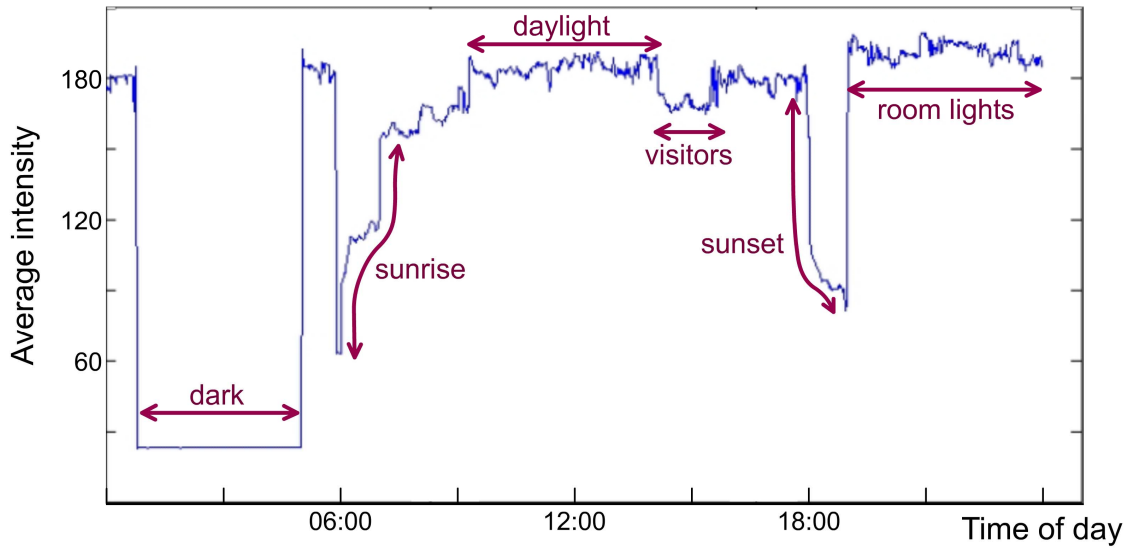


Figure 18: Lighting changes in the living room and the corresponding events.

7.2 Event detection based on lighting changes

We intend to use change of light level as a cue for video retrieval from all the locations in the house. Lighting changes in a room can take place due to very simple events, such as turning a light on/off or opening a window curtain. However, if combined with the scene context, they can be used to identify significant events that take place in a house. For instance, lighting changes in the rooms at night will provide a rough idea as to when the residents went to sleep, and whether some of them woke up at night etc. Being very quick events, they can be used to create very short summaries of a day's events. Figure 18 shows the variation of light level inside the living room of ubiquitous home, during a full day. The text labels indicate the actual events that caused the corresponding light levels and changes.

We presume that the lighting level in a selected region of ubiquitous home at a given time can be represented by the average intensity of all pixels in all the images captured in that region at that time. Lighting changes are relatively easy to detect, as

they are represented by sharp changes in average intensity calculated as above. However, the problem is to find a threshold level for this change that is suitable for separating significant events (such as entering a room in the morning when it is partially lit from outside) from insignificant events such as a curtain being blown away by the wind. For rooms or regions with windows, the amount of external light changes with the seasons, weather, curtains, and time of day. For ubiquitous home, this tasks is made further complicated by automatic gain control in the cameras. Setting a single threshold level to match all these conditions is impossible.

Our approach to solve this problem is to assign a rank of significance to each lighting change, based on the sharpness of change. The user selects the rank and browses the events through an interactive interface, thereby reducing the search space intuitively.

We consider a day as the unit of video to be processed simultaneously, as it includes a full cycle of lighting variations both inside and outside a house. For each camera, the frame intensity function $I(t)$, where $I(t)$ = the average intensity of the frame t, is calculated. This is low pass filtered by averaging over a window of 5 frames (corresponds to a 1 s interval).

The intensity gradient function, $g(t)$ is calculated as

$$g(t) = \left| I(t+1) - I(t) \right|$$

The small intensity variations, which are mainly due to moving objects and persons, are removed by thresholding $g(t)$. For this, the threshold for each camera was

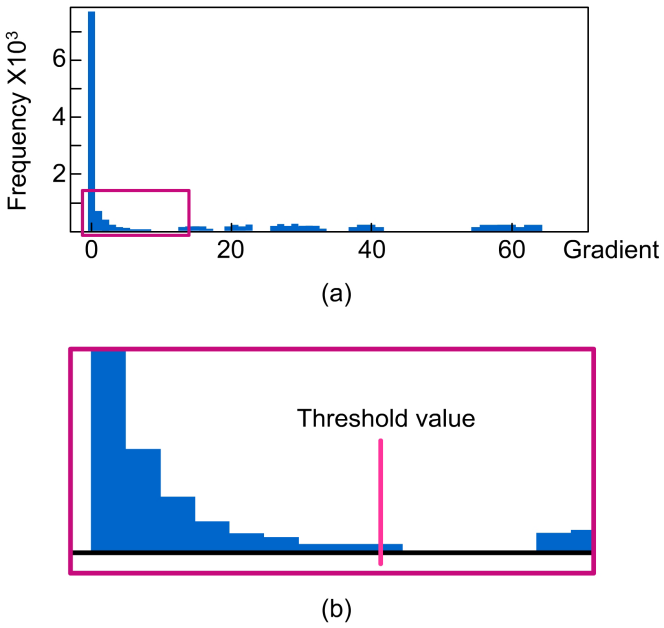


Figure 19: Threshold estimation using gradient histograms.

determined using 24 hours of training data. The intensity gradient function is calculated for the training data set and its histogram with a bin size of 1.0 is constructed (Figure 19a). The half-Gaussian shaped cluster of bins at the beginning of the histogram corresponds to insignificant lighting changes. The threshold value H is selected at the last bin value in this cluster. Figure 19b shows an enlarged section of the histogram in Figure 19a with the threshold value detected using the above method.

The thresholded gradient function, $G(t)$ is defined as

$$G(t) = 0 \text{ if } g(t) < H$$

$$G(t) = g(t) \text{ otherwise}$$

The scaled gradient function, $N(t)$ is determined by

$$N(t) = 0 \text{ if } G(t) = 0$$

$$N(t) = 1 + \frac{G(t) - G_{\min}}{G_{\max} - G_{\min}} \times 9 \quad \text{if } G(t) > 0$$

Where $G_{\min}$ and $G_{\max}$ are the minimum and maximum gradients recorded during the selected date.

The objective of scaling is to set the range of $N(t)$ to [0 10], with all lighting changes scaled to the interval [1 10]. This will give the same priority to events taking place in different regions in different lighting conditions. In case real-time calculation is necessary, $G_{\min}$ and $G_{\max}$ can be estimated using previous data, although this will cause a minor change in the range of $N(t)$.

The rank function $R(t)$ for each region is determined by taking the summation of $N(t)$ for all the cameras in that region. The partitioning of regions is the same as shown in Section 4.4. It is evident that $R(t)$ is higher for sharper lighting changes, and for lighting changes seen by all cameras in the region. The user, at the time of retrieval, can interactively specify $R(t)$ and retrieve events. This is described in detail in Chapter 8.

7.3 Action Classification for Retrieval

Step segmentation and video handover results in video and key frame sequences. However, these can be lengthy if the persons tracked stayed a long time in the house. Furthermore, it is desirable to partition these results further according to the actions they performed.

7.3.1. Clustering of Footstep Sequences

We observed the results of clustering different combinations of variables in sensor

activation data, using Kohonen Self Organizing Maps (SOM). The activation durations showed a grouping that is independent from other variables. Durations between 0.10 and 0.96 seconds formed a distinct cluster consisting of 90% of the data. To examine if this grouping leads to any meaningful summarization, the video data was retrieved using the following approach. Sensor activation data was segmented using [0.10, 0.96] seconds as the threshold interval. The activations that occur with less than 1 second time gap in between were clustered to obtain activation sequences, corresponding to time intervals. Video clips for these time intervals were retrieved from the relevant cameras and examined.

It was evident that video clips corresponding to the segment with durations > 0.96 s corresponded to video containing activities with irregular or infrequent foot movement, such as sitting, waiting, and preparing food. The rest corresponded to walking and vacuum cleaning. Therefore, clustering using this approach enables retrieval of short video clips pertaining to two basic categories of actions.

7.3.2. Detailed Action Classification

Clustering of sensor activations, as described in the previous section, results in only a basic classification of activities. It is necessary to have more specific activity classification, to be able to retrieve video for queries related to daily life. An action database (Table 15) was constructed by extracting portions of footstep sequences created in footstep segmentation of the data from the real-life experiment. The selected actions seem somewhat unbalanced; for instance, walking and standing are basic body gestures and cooking is an activity with higher detail. However, the selected actions are those appearing most frequently in footstep sequences retrieved from ubiquitous home.

Therefore, it is more practical to train a classifier for these actions. The duration of the sequences was between 30 seconds to 5 minutes. Each sequence contained a minimum of 20 sensor activations.

Sensor activations taken individually do not reflect the dependence of the footsteps on previous footsteps and the relationship within a group of footsteps. For example, a standing person will keep the feet at nearly the same place, with occasional changing of weight on one foot to the other whereas a walking person has only one foot on the floor most of the time and keeps a somewhat regular distance between consecutive footsteps. We define and calculate the following features for activity classification for an activation sequence $S = \{A_1, A_2, ..A_n\}$ where A_i is the i^{th} sensor activation of the sequence.

- Mean and standard deviation of sensor activation durations
- Standard deviation of X coordinates of the sequence
- Standard deviation of Y coordinates of the sequence
- Overlap of footsteps O_s, defined as

Table 15: Composition of the activity database

Action	No. of sequences
Walking	40
Standing	56
Sitting on a chair	30
Sitting on the floor	10
Cooking or washing dishes	22
Vacuum cleaning	10

$$O_s = D / T$$

Where

D = sum of durations in all activations $\{A_1, A_2, ...A_n\}$

and

T = Duration of A_n + start time of A_n - start time of A_1

- Activation rate R, defined as

$$R = n / T$$

The sequences in the activity database were classified and tested using WEKA Machine learning tools [68]. Multi-layer Perceptron (MLP), *k*-Nearest Neighbor and Random Forest classifiers with different parameters were trained and tested on the database. The results were evaluated using 10-fold cross validation. An MLP classifier with 3 layers yielded the highest overall accuracy, and therefore was selected to be implemented in the proposed system.

7.3.3. Combining other modalities to improve accuracy

The data from other sensors were used to improve the accuracy of activity classification, based on the following heuristic rules:

- For vacuum cleaning, audio classification for the same region should detect vacuum cleaning sounds for the corresponding time interval. Otherwise, the action is re-classified as walking.

- A cooking event is rejected if at least 60% the X and Y coordinates of the sequence of footsteps are outside the area near the kitchen pantry. In this case, the action is re-classified as standing
- For sequences classified as “sitting on chair”, the X and Y coordinates should lie within predefined regions near the chairs. This has a problem in the long term as people tend to rearrange furniture at times, but is logical for heavy chairs like sofas for a small house and duration in the order of months.

7.4 Evaluation

7.4.1 Event detection based on lighting changes

The number of lighting change events detected per hour was in the range of 0 to 8. It was possible to detect fine changes such as opening a door and entering a room that is already lit and sharp events such as turning lights of a room on/off at night. Due to the positioning of cameras and automatic gain control characteristics, very few false positives caused by moving people close to the camera were detected. Figure 20 shows the lighting changes that took place in the living room between 5:00 a.m. and 6:00 a.m. on the 12th of April 2005. For each event, images captured before and after the lighting change from one camera are displayed. By looking at only these two events, it is possible to understand that the family members in the photo entered the living room at 05:01a.m., and one of them left the room at 05:54 a.m.

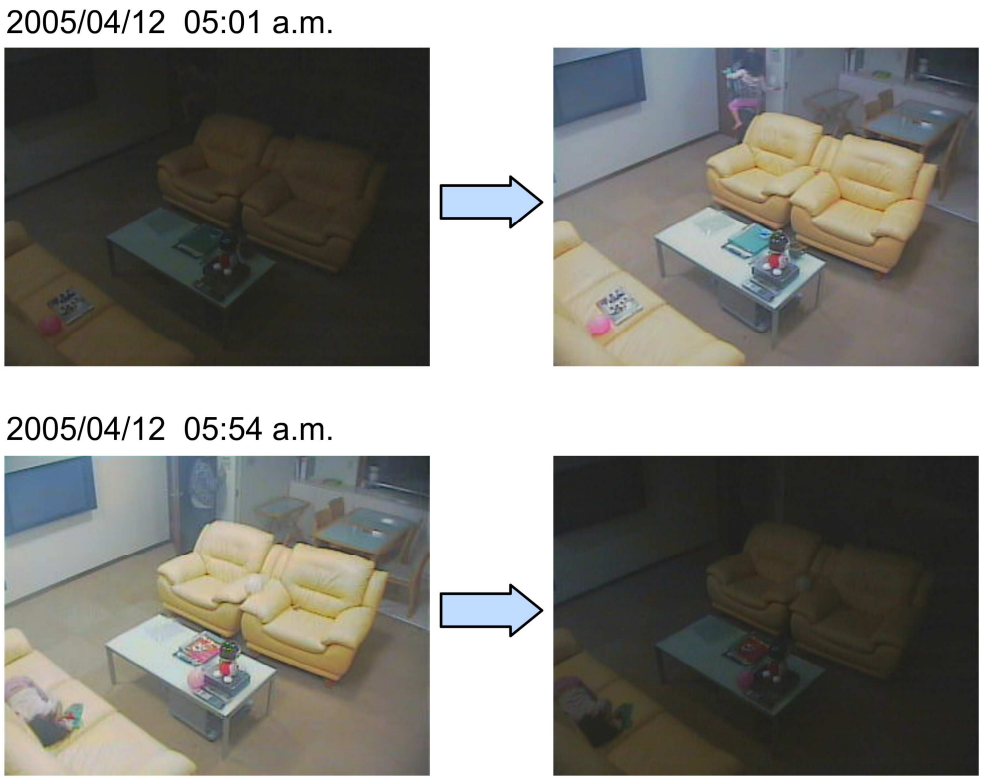


Figure 20: Retrieved events for lighting changes.

7.4.2 Basic Activity Classification

We calculate precision P, recall R, and balanced F-measure F for evaluation of retrieval of video with regular foot movement.

$$P = N_c / (N_c + N_m)$$

$$R = N_c / (N_c + N_o)$$

$$F = 2PR / (P + R)$$

Here N_c is the number of correctly retrieved video clips, N_m is the number of clips that were not retrieved, and N_o is the number of mistakenly retrieved clips. Step

sequences with accurate step segmentation were clustered to retrieve video clips and the clips observed to evaluate the performance. The precision of retrieval was 93.7% and the recall 96.7 %. The F-measure was 95.2%.

7.4.3 Detailed Action Classification

Table 16 presents the results of action classification for the selected classifier. It is evident that the accuracy is lower for recognizing the two types of sitting actions, and vacuum cleaning. An interesting observation is that cooking can be distinguished from standing with a high accuracy, despite using only floor sensor data and even without using the location of the person as a feature. Further examination of data revealed that

Table 16: Accuracy of action recognition before using multiple modalities.

Label	Action	Precision	Recall	F-measure
1	Walking	0.848	0.780	0.813
2	Standing	0.811	0.750	0.779
3	Sitting on a chair	0.455	0.625	0.526
4	Sitting on the floor	0.571	0.500	0.533
5	Cooking or washing dishes	0.902	0.920	0.911
6	Vacuum cleaning	0.600	0.714	0.652

Table 17: Confusion matrix before using multiple modalities.

Correct action	Classified as					
	1	2	3	4	5	6
1	39	1	1	0	0	9
2	1	30	2	1	5	1
3	0	1	5	2	0	0
4	0	1	3	4	0	0
5	0	4	0	0	46	0
6	6	0	0	0	0	15

there is higher overlap of footsteps and activation ratio for cooking compared to standing. While cooking, a person tends to make a number of small foot movements and puts weight on both feet due to the need to balance the body while handling the kitchen appliances. A person who is just standing, on the other hand, usually rests his/her weight on one foot and switches it regularly, rather than moving feet. These differences were represented well in the selected features, allowing classification with high accuracy.

The confusion matrix (Table 17) shows that most of the confusions occur between the two types of sitting actions. Walking and vacuum cleaning is another pair with a large amount of confusion. The heuristic rules involving other modalities were selected considering the patterns of confusion. Tables 18 and 19 show the accuracy and the confusion matrix after applying these rules. It is evident that there is a 1% to 18% improvement in the overall accuracy represented by F-measure.

7.5 Discussion

Given the large amount of information contained in image data, it should be possible to detect more actions and events using image analysis. However, we restricted our work on image analysis for the ubiquitous home, owing to the following reasons. Due to the constraints with disk space, they have a low frame rate (5 frames/second), low resolution (320x240), and a high degree of lossy (JPEG) compression with poor quality (15 kB/image). Automatic gain control is used so that there is maximum possible visibility in the captured images. Such images are sufficient for human observation upon retrieval, but not adequate for obtaining high accuracy using existing image

Table 18: Accuracy of action recognition after using multiple modalities.

Label	Action	Precision	Recall	F-measure
1	Walking	0.873	0.960	0.914
2	Standing	0.821	0.821	0.821
3	Sitting on a chair	0.500	0.625	0.556
4	Sitting on the floor	0.625	0.625	0.625
5	Cooking or washing dishes	0.939	0.920	0.929
6	Vacuum cleaning	1.000	0.714	0.833

Table 19: Confusion matrix after using multiple modalities.

Correct action	Classified as					
	1	2	3	4	5	6
1	48	1	1	0	0	0
2	1	32	2	1	3	0
3	0	1	5	2	0	0
4	0	1	2	5	0	0
5	0	4	0	0	46	0
6	6	0	0	0	0	15

analysis algorithms. For example, face recognition is impossible given the resolution and compression. A pilot study on face detection using Viola-Jones feature detector [69] yielded a maximum accuracy of 86.5% [70]. However, this required scaling and smoothing of images, which need resources in terms of both processing power and disk space.

The actions and events that are detected using the above algorithms are fairly basic, and there is room for further work in action and event recognition. However, we stop at the current state, demonstrating the applicability of data driven selection of actions and features, supervised learning and sensor fusion for accurate retrieval.

Chapter 8

User Interaction

8.1 Issues

The user retrieves video, audio and key frames through a graphical user interface. The main purpose of a user interface is to enable the users to accept queries from the users and present the results to them in a comprehensible manner. However, it is evident that there is a considerable difference between the semantic levels of the results obtained in Section 4 and the user queries listed in Section 1. Commonly known as "semantic gap" [71], this causes problems in most multimedia retrieval systems. If the gap is closed by making the users submit lower level queries, the usability of the system goes down. On the other hand, trying to fill the gap using heuristics or simple assumptions will result in lower accuracy. Our approach to solve this problem is twofold. From the side of the system, better visualization of events is provided using *hierarchical media segmentation*. At the same time, user intelligence is incorporated to the query process by means of interactive queries. The following subsections describe these two main concepts and the design of the user interaction in detail.

8.2 Approach

The user interaction was designed in two main stages while the research was in progress. First, a simple graphical user interface was designed to access video by selecting a date, time interval and a particular camera. More functionality added to this interface with the implementation of the algorithms in the previous chapters, using evolutionary

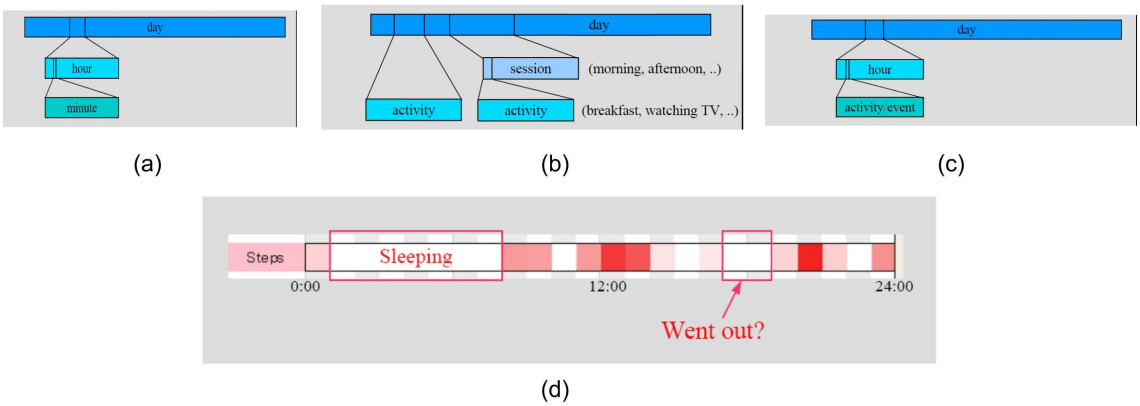


Figure 21: Hierarchical media segmentation.

prototyping technique [72]. After developing a prototype version that was used for a user study with residences (Chapter 9), the user interaction strategy with system was redesigned according to user feedback and based on the concepts described in the following sections.

8.3 Hierarchical Media Segmentation

The system captures multimedia data continuously, from which segments corresponding to events are retrieved as results. The original data from ubiquitous home are indexed by source (camera ID, microphone ID, floor sensors coordinates) and timestamp. The results obtained using the algorithms in Section 4 are indexed by the timestamp, location and event/action. Figure 21a presents the segmentation of data after capture and analysis.

However, these indices are still unable to facilitate efficient multimedia retrieval, as humans tend to remember the events and would like query for events that are segments in a different way. For example, “Retrieve video showing the regions house where people were in at 9:47 a.m.” will be a valid query for the system, but is very unlikely for a human user to enter. A more likely query is “What was I doing after

breakfast last Saturday?". Humans tend to segment the time and experiences at home by days, sessions (e.g.: morning, late afternoon), locations, and events (Figure 21b). However, these are difficult to model using a computer-based system, due to high semantic level and fuzzy boundaries of the segments. We propose a trade-off between these two levels, shown by Figure 21c. The media is segmented hierarchically by date, hour, location and event. Visualization of a daily summary allows the users to identify sessions. For example, the user can identify the sessions of the day using the summary of floor sensor activity, sound and lighting level, as shown in Figure 21d. This method is an extension of the concept of *hierarchical timeline segmentation* [73].

8.4 Interactive Retrieval

The system, at its current state and with the available sensor data, is unable to perform some useful tasks (e.g. person recognition). Furthermore, the accuracy for the algorithms that are implemented is less than 100%. However, the performance can be improved greatly if it is possible to incorporate user's intelligence to the system. We propose interactive retrieval to achieve this. A query is broken down to a number of steps, and each step returns intermediate results to the user so that the user can provide further input navigating towards the desired results. For example, if the user wants to retrieve video showing what person *A* did during a given time interval, the system provides a key frame from each video sequence created by segmenting footsteps. The user can take a look at the key frames and identify those showing person *A*, resulting in accurate retrieval in the expense of one additional step.

8.5 User Interface Design

The user interface has been designed with the following objectives:

1. Ability to use with only a pointing device (either a tablet monitor or a touch monitor).
2. Require a minimum technical knowledge for understanding the inputs and results: all user-adjustable parameters will be interpreted to the user in a way they understand. For example, a slider control input labeled as “Sampling rate gradient for key frames” with range 0.0 to 1.0, has little meaning in the user’s perspective. This can be modified by labeling the input as “Desired amount of key frames”, with *few* and *many* labeling the ends of the slider.
3. Facilitate easy navigation within data without starting over: humans tend to search for *relative queries* such as “what happened next?” and “what happened in the other room during this time?”. We attempt to facilitate such queries by dividing the results into a set of tabs and update the results in tabs according to the parameters submitted for the last query. While browsing the results, a user can navigate along the timeline outside the query boundaries, using button-based inputs such as *previous* and *next*.

The user interface is designed to have more intuitive inputs. For example, the users can click on an image showing the home layout, to select a room/region to retrieve events from. Camera selection is facilitated in the same manner, and the view from a camera is immediately available to the user so that the right camera can be selected

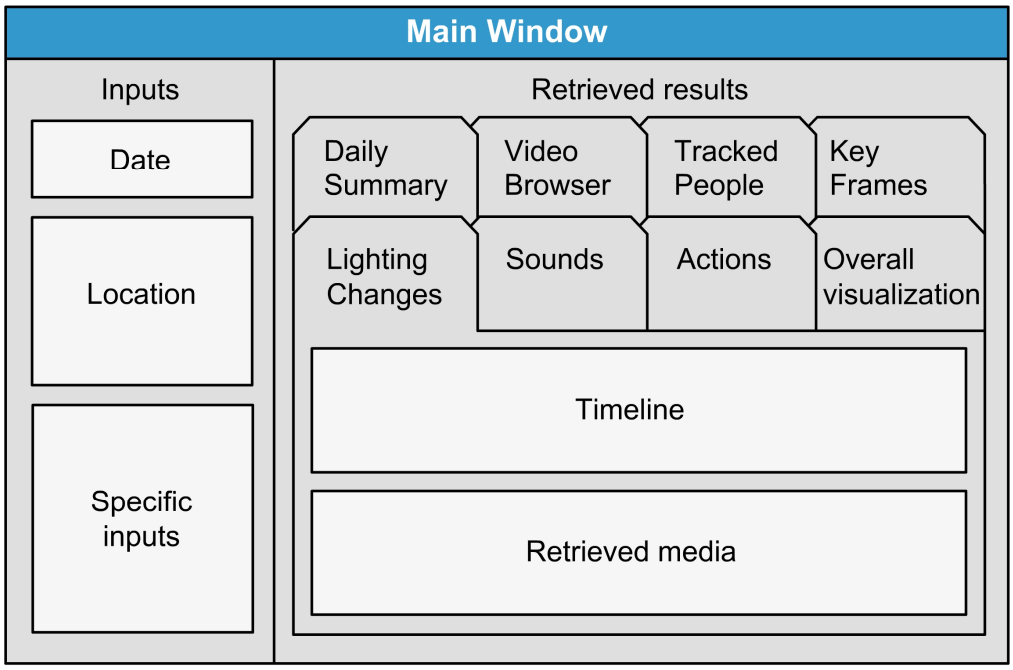


Figure 22: Organization of the user interface.

interactively. Instead of line or bar graphs, color levels are used to indicate the level of each activity as this is faster and easier to interpret.

8.6 Presentation and Visualization of results

Figure 22 shows the basic organization of the interface. The main window is divided into two main sections, as inputs and results. The input section includes the user inputs for date, location, and other specific options such as time intervals. The results for queries based on inputs on the left are grouped into tabs, as shown enlarged in Figure 22. All the tabs are updated at the same time when a user changes inputs, so it is possible to see different groups of results by merely clicking on the appropriate tab. The following subsections describe the presentation of each of these groups.

8.6.1 Daily summary

The user starts by entering a day using a calendar interface, upon which a summary of the day's activity is displayed along the time line (Figure 23). Stripes of different colors, segmented in to one hour intervals, are used to represent different types of data captured and results retrieved during the selected date. On each graph, the strength of the corresponding color, indicates the amount of data/results present. White corresponds to no data/results, whereas the strongest color saturation indicates the maximum amount of data present during that day. Numbers are deliberately excluded to avoid information overloading. The colors used in all visualizations are scaled so that the results are shown to the user with maximum possible contrast. An example situation where this is useful is when all the lighting changes are less sharp during mid-day. Due to scaling of colors for visualization, the user is still able to see them clearly. The user can select the location of the house to retrieve the summary from, if necessary, using the inputs on the left.

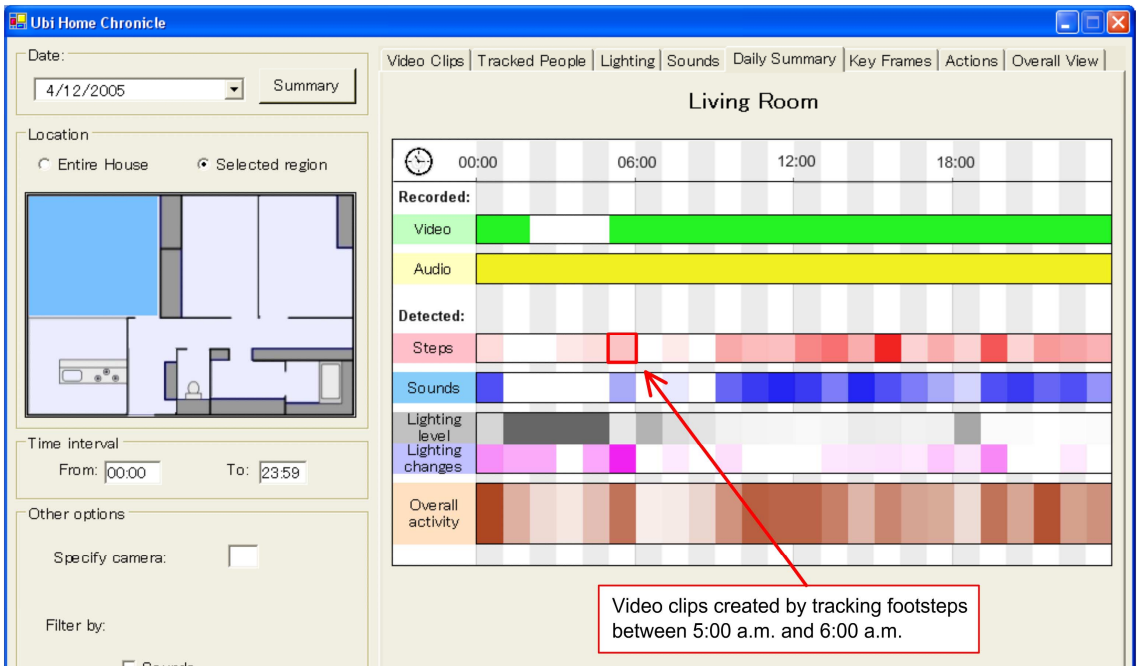


Figure 23: Visualization of the daily summary.

Figure 23 is a screenshot of the daily summary for the living room, on the12th of April 2005. It is possible to interpret a fair amount of activity by simply observing the daily summary. It is evident that video has not been captured between 2:00a.m. and 5:00 a.m. on this date. Looking at the graphs labeled "Steps" and "Sounds", it can be deduced that the residents have slept after midnight and before 1:00 a.m. It can also be seen that a large number of footsteps in the living room have been detected between 3:00 p.m. and 4:00 p.m. This can be due to having visitors (in this case the actual reason), children playing, or a special event such as a party.

The user can click on any one-hour block on the daily summary, to see the selected type of results for that one hour. For example, clicking on the block marked with a red border in Figure 23 will show a summary of media for "tracked people" in the house between 5:00 a.m. and 6:00 a.m., which can be browsed further to see the

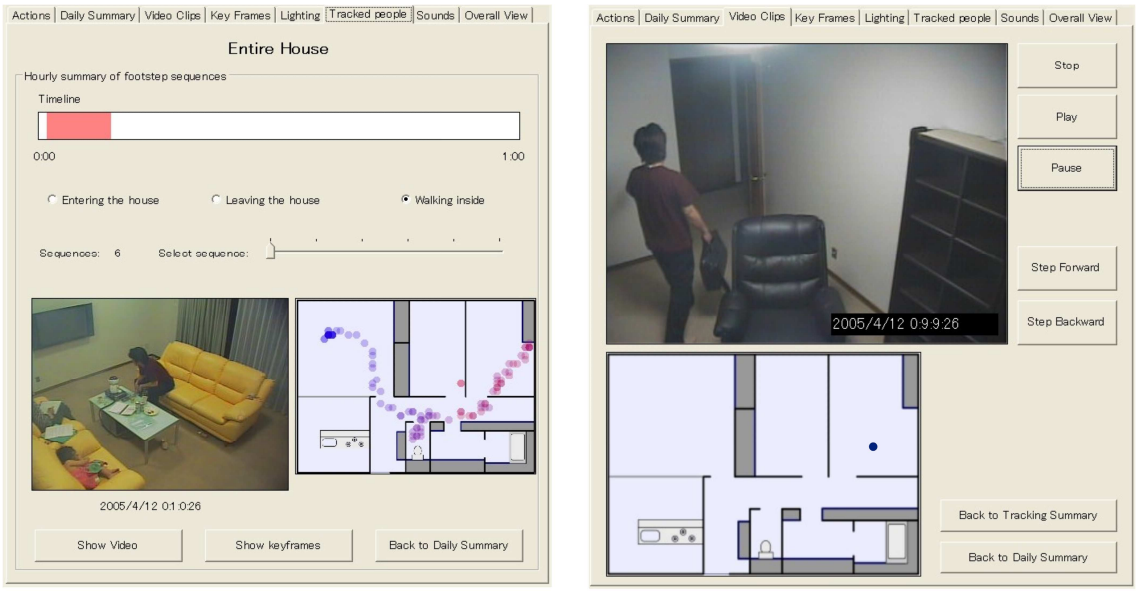


Figure 24: Viewing video for tracked people.

results from personalized video retrieval. The following subsections describe this and similar options that are available to the user.

8.6.2 Tracked people

Figure 24a shows the retrieval of footstep sequences for a selected duration of one hour. The timeline now shows only this duration. The User can select footstep sequences corresponding to instances of persons entering the house, leaving the house or walking inside. The resulting sequences can be previewed one at a time, using the slider control. The timeline shows the duration of the selected sequence. This can be modified further to indicate the type of action the person is performing, using different colors. The preview image shows the first frame of the video clip created for this sequence. The dots on the house floor plan show the path of the person's footsteps. The color of the dots changes from blue to red with time, indicating the direction of the person. Thus the

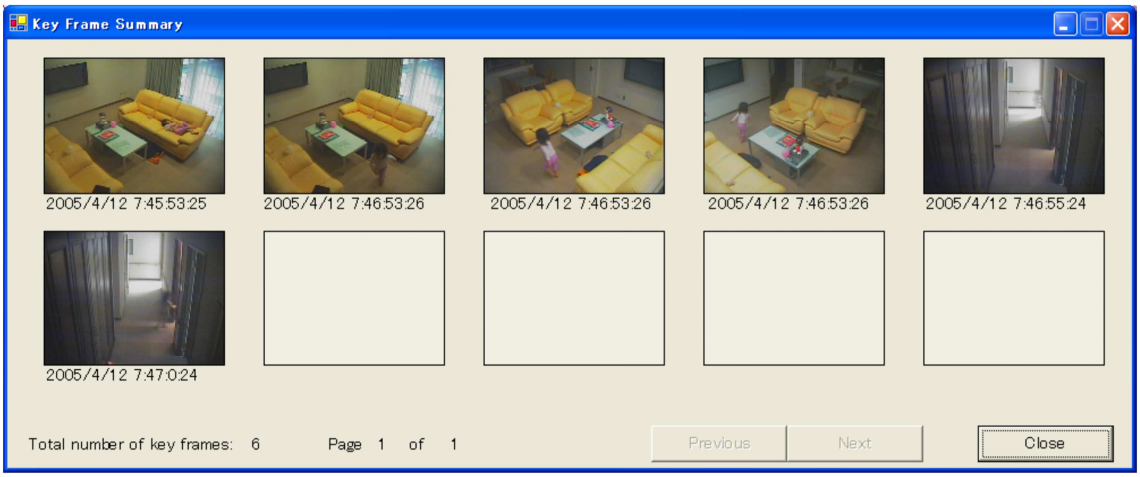


Figure 25: Displaying key frame sets.

user can interpret the video sequence to a certain extent even before fully viewing it, and thereby find the desired results faster. After selecting the desired video clip, the user can play it at normal speed, or browse it using the video clip viewer in Figure 24b. A moving dot on the house plan will now show the location of the person.

8.6.3 Key frames

The user can choose to view a sequence of key frames after selecting a footstep sequence from the preview window describes in Section 7.6.2. The key frames extracted using the adaptive spatio temporal sampling algorithm are shown to the user (Figure 25).

8.6.4 Sounds

The main elements of the preview window for video retrieved for sounds are the same as for the footstep sequences. The duration of each sound segment or class is shown along the timeline when the user selects it (Figure 26). After selecting the desired sound clip, the user can retrieve video from the cameras in the region where the sound was

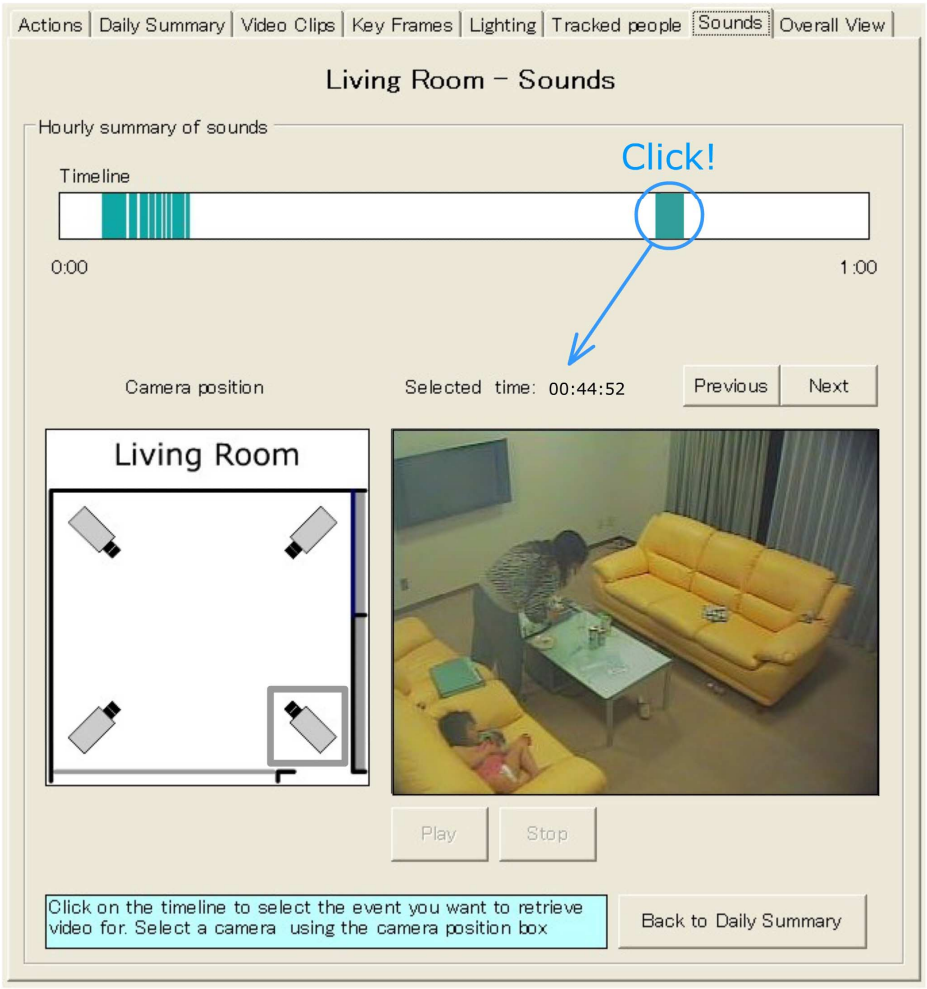
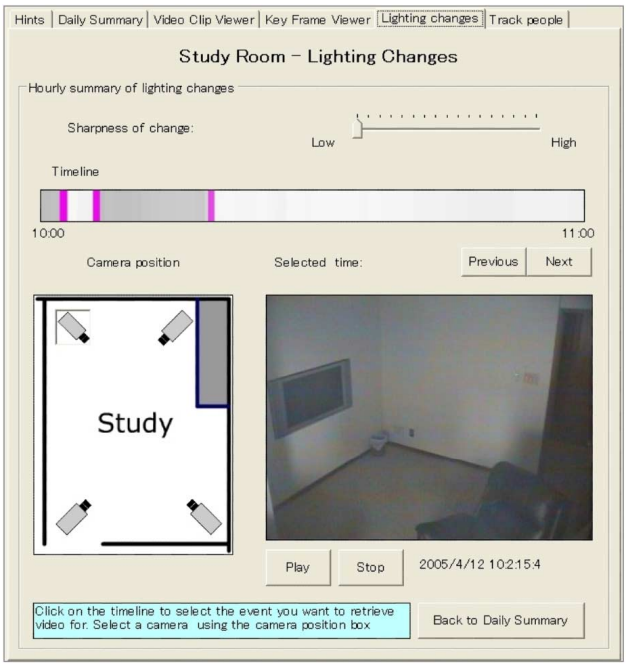
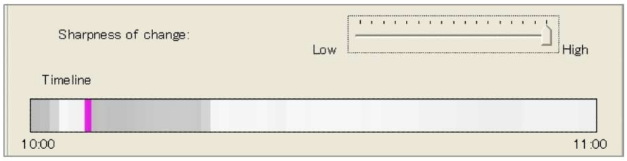


Figure 26: Retrieving video for sound segments.

generated. The desired camera can be selected interactively while watching the video, by clicking on the appropriate camera drawn on the floor plan of the region.



(a)



(b)

Figure 27: Interactive retrieval of lighting change events.

8.6.5 Lighting change events

Figure 27 illustrates the use of interactive retrieval to retrieve video for events that cause lighting changes. The light level in the region, scaled over the given date, is shown on the time line so that the lighting changes can be interpreted easily. The slider labeled as “sharpness of change” is coupled to a threshold within the range of the rank function. When this set to minimum, all events are displayed on the time-line (Figure 27a). When set to maximum, only the event that has the maximum rank during the selected one hour time interval is displayed (Figure 27b). Selection of the camera for

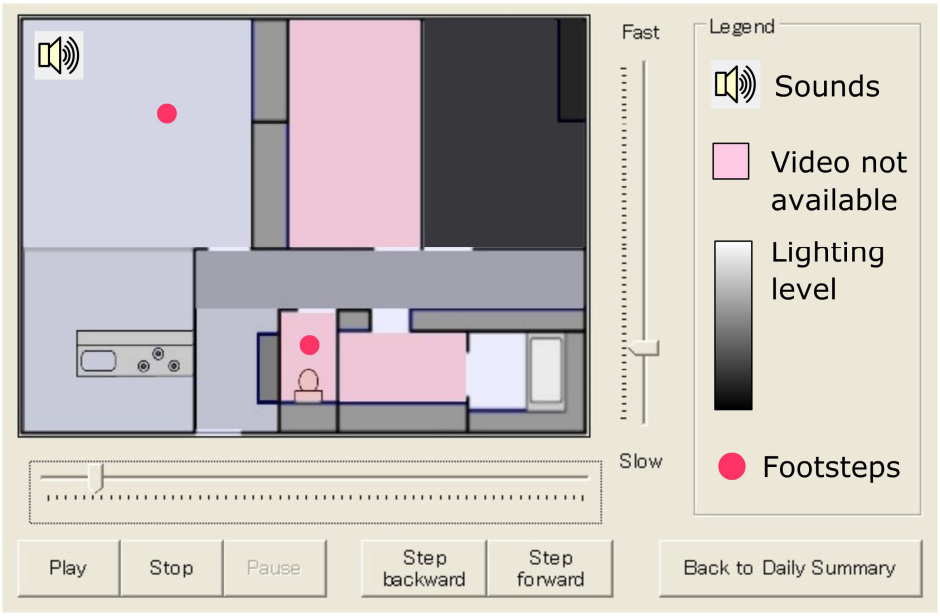


Figure 28. Animated preview of overall activity.

viewing is done interactively by clicking on the image of the camera on the room layout, thereby providing an intuitive method to select the viewpoint (Figure 27a).

8.6.6 Overall activity visualization

The interface components describe above display results of one category, most of the time from one location. However, it is desirable to have a simultaneous overview of what is happening in the entire house, if possible. The main problem in achieving this is finding a method to visualize the various types of data from a large number of sources, without overloading the user with the data. For instance, there are 17 cameras in ubiquitous home, recording data continuously. If the videos from these cameras are displayed concurrently, it will be extremely difficult for the user to focus on everything and get a general idea of what happened. The same applies for audio data, too.

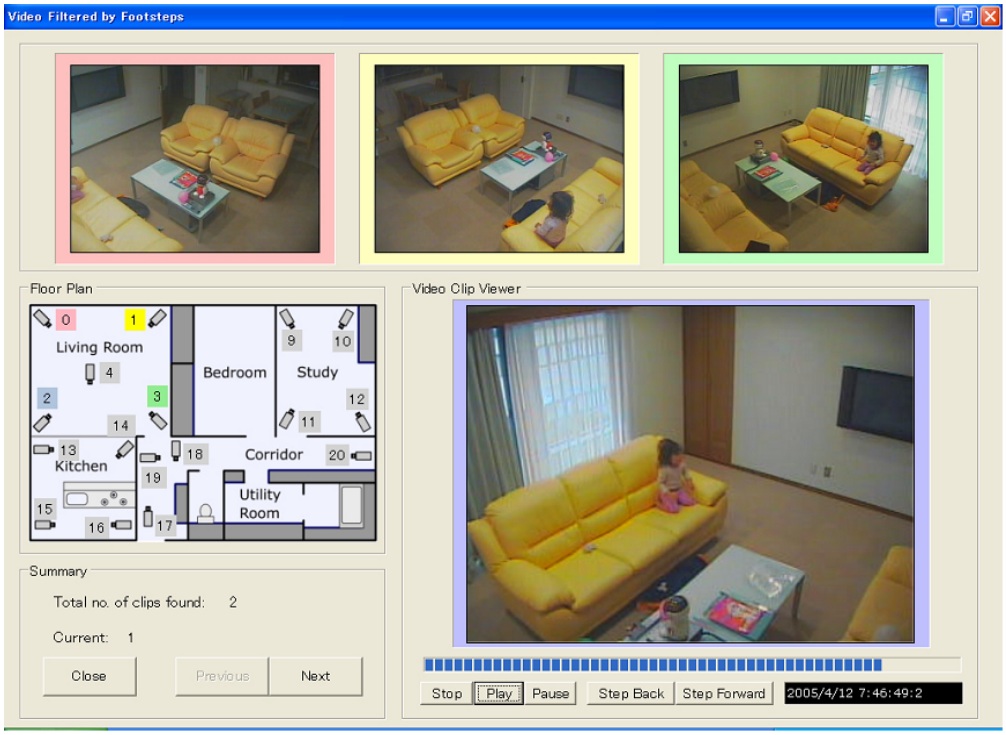


Figure 29: Video browser for multi-camera preview.

We propose a simultaneous visualization of the desired results, to facilitate an overall visualization of activity inside ubiquitous home. Figure 28 shows a screenshot of this visualization. The proposed visualization is an animation with a timestamp, where symbols are displayed plan of the house to visualize different types of activity. The brightness of each region in the plan corresponds to the actual light level in that region of the house. Dots are displayed to represent floor sensor activations. A speaker is displayed in each region when there is a local sound segment from that region. The speed of the animation can be changed so that it can be browsed faster to see what happened.

8.6.7 Video browser

The video browser is an extension of the simple form of a multi-camera surveillance system. The user can select a location and a time interval to retrieve video from. It is also possible to select a *main camera* for retrieving images. The video from this camera will be shown larger whereas video from other cameras in the same region will be shown in smaller size. Unlike in a normal video surveillance system, the resulting video can be filtered to retrieve video for situations only with sound and/or footsteps. Figure 29 is a screenshot of the video browser. The colored border around each picture frame indicates the camera it is retrieved from.

8.7 Example Scenario of Retrieval

Suppose Mrs. Taro wants to see what her son, Takeshi, did after coming back from school on a particular date and before leaving for sports practices in the evening. Also, suppose she remembers that Takeshi came back home house some time after 2:00 p.m. and left home around 3:00 p.m. For a conventional retrieval system based on date, time and camera, it is necessary to watch the video from the camera showing the entrance to the house from 2:00 p.m., 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. This becomes increasingly tedious with movement and the duration of Takeshi's stay.

The proposed system can be used for this query, in a much simpler manner. First, Mrs. Taro enters the particular date and retrieves the daily summary. On the "Footsteps" graph of the summary, she can click on the one-hour block corresponding to the duration between 2:00 p.m. and 3:00 p.m. After selecting "Entering the house", she can

see the set of preview frames showing people who entered the house during this time interval. By browsing these preview frames showing the persons entering the house, Mrs. Sato can find the key frame showing Takeshi. She can now see either the complete video clip showing what Takeshi did inside the house. Since the cameras and microphones are selected automatically, no further user input is required. If necessary, Mrs. Sato can view a set of key frames instead of the complete video, achieving even faster retrieval.

8.8 Discussion

Designing user interaction and developing a graphical user interface is a challenging and iterative task, as the user feedback often includes requests for increased functionality. While a badly designed interface requires changes, a good and usable interface stimulates imagination of the users and results in requests for added functionality. Therefore, it is essential to carry out evaluations regularly while adding new retrieval algorithms to the system. However, for the ubiquitous home, it is quite difficult to use residents as subjects for user studies regularly. Therefore we conducted a single, extensive user study with residents and used the results for improving the system.

The results from some of the stages of data analysis are yet to be integrated to the current version of the user interface. This is mainly due to the parallel development of both the user interface and the algorithms for data analysis and multimedia retrieval. However, this could be completed by merely adding database queries and program modules to display the results, without any research effort. Furthermore, the design of user interaction is sufficiently modular so that new types of queries and visualizations can be added to the interface without changes to the current system.

The system could benefit from using different input devices and output visualizations. In a 5-day demonstration session with a large number of users, we found that users found the system more natural to browse using a graphic pen tablet monitor, instead of the conventional keyboard and mouse. While the current two-dimensional visualizations are easy to use and informative, three-dimensional visualizations with improved navigation capabilities are worth investigating, due to the large content and three-dimensional nature of the actual environment.

Chapter 9

User Study

While the evaluation of individual stages show that each stage has an accuracy between 60% and 98%, it is still essential to conduct a user study where the user retrieve their own experiences in a home environment. We conducted such a user study with a family who stayed in ubiquitous home during one of the real-life experiments. The coming sections of this chapter describe this user study and the results.

9.1 Objectives

The user study was designed with the following objectives:

- Identify requirements for experience retrieval in the ubiquitous home, using feedback from people who lived in the environment.
- Evaluate algorithms that have been implemented so far for video summarization and retrieval.
- Evaluate the usability of the system as a multimedia mining tool for use by residents of a household with a non-technical background.
- Identify directions for future work and improvement of the existing algorithms.

Since there was no previously designed evaluation experiment that fulfills the above objectives, we designed a detailed user study that is specific to our system. Since the system was being developed at the time the user study was conducted, we used a prototype system that did not have some of the functionality described in Chapters 6

and 7. However, this was not a problem or weakness as one of the objectives of the user study was to identify the user requirements, and the feedback received could be utilized for implementing a better system.

9.2 Participants

We selected a family who stayed in ubiquitous home for two weeks, during one of the real-life experiments. The family had three members; a married couple, and their 3-year old daughter. On the working days, the husband went to work and the wife stayed most of the time at home, taking care of the child. The family went out occasionally for meals, shopping and other family activities, but returned every night as usual for an ordinary household. There were a few guests, but they did not stay overnight. The subjects were paid for participation in the user study.

9.3 Procedure

This study consisted of two parts; a requirements analysis for home experience retrieval, and a hands-on session of the system we developed. Data captured during six hours on the 12th of April 2005 (the 4th day of the family's stay in ubiquitous home) were used for the study. This was equivalent to retrieving from 102 hours of video and 150 hours of audio data. The date of the user study was the 12th of October 2005, exactly 6 months later.

The experiment consisted of three sections. The first section was a requirement study, where the subjects answered a questionnaire to specify what they would expect from a system for retrieving experiences at home. This section of the experiment was

conducted before demonstrating the system, to ensure that the user requirements are not influenced by the functionality of the existing system.

In the second section, the subjects were given a demonstration on how to use the system. Only one example for each type of retrieval was shown. Thereafter, they were allowed to use the system themselves, submitting their own queries to retrieve their experiences. The authors were available in case the subjects needed advice, but were not involved in using the system. The subjects were asked to select video clips that they would like to keep, so that we can provide them in a DVD. This was done both as a factor of motivation and also to find out what kind of experiences generate interest in keeping a permanent record. After using the system, the subjects rated the usability of the system by answering a brief questionnaire based on the guidelines by Chin et al.[74].

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

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

9.4 Results

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

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

The subjects added the following after using the system.

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

2. Supposing it is possible to have the same facility at your home, and only your family has access to the data:

    (a) What would you like to use it for?

    - For taking care of my old mother; check whether she took medicine properly, or she ate too much sweets, etc. (1)
    - For finding lost objects, discover our own habits, find out how the child is behaving so that anything bad can be corrected (1)

(b) How would you like to record the data?

- Record everything from daily life (2)

(c) Which parts of the house will you record everything?

- Non-private areas of the house, like the living room, the kitchen, etc. (2)
- Child's room (1)

(d) Which parts of the house and times will you refrain from recording?

- Private places, such as bedrooms (2)

(e) Do you have any other preferences, such as times of day, about recording?

- Would like to see my child during day time and afternoon, when I am not at home(1)
- Want to record leisurely times playing with the child (1)
- Would like to record busy hours of the day for discovering things that were left behind (1)

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

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

- Overall impression – 5,6

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

1. How much did you remember from what you could see in the video and key frames?
    - There were many things that I did not remember. For example, that I worked that day (1)
    - I roughly remembered what happened on that day. But the memory was refreshed a lot after watching the video (1)

2. Was it possible to see interesting things that you did not see/know before?
    - Yes (2)
    - We could discover things like how our child woke up in the morning (1)
    - I was surprised to learn that I spend so much time with my child (1)

3. Out of what you saw, which parts of the video would you like to keep with you?
    - Video of the child (2)
    - Video of activities we did together, such as having meals (1)

4. State what you like about this software.
    - Automatic camera change (1)
    - Ability to see what happened when I was away (1)

5. State what you don't like about it.
    - It might reveal things that are not nice to know (1)
    - Too many video clips and key frames to look at (1)

6. For what kind of things will this software be useful to you?
    - Family diary, security
    - Taking care of family members
    - Record of our child's life
    - See myself objectively.

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

The subjects managed to recall what happened in the entire session and to retrieve video they wanted to watch, by using the system. They found two types of video more interesting, and watched them repeatedly. One type contained video of the child when she was alone: an example was the video clip created when the child woke up in the morning, found that she was alone in the living room, and ran for the mother. The other type corresponded to activities that they did together, such as taking meals and playing with the child. They requested copies of both these types of video clips. The subjects used key frames as an index to the original video, rather than viewing only the key

frames as a summary. They liked using the system, and it was somewhat difficult to get them to stop watching videos and answer the questionnaire.

9.5 Discussion

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

With only two persons actively taking part in the user study, the responses have little statistical value. However, their keen interest on using the system and positive feedback justifies the motivation and the current progress of this work. The responses also provide valuable insights to identify further requirements and possible improvements. Continuing further study with other families, as the system is being developed, will help the system to evolve into one that is very useful.

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

Chapter 10

Conclusion and Future Work

10.1 Conclusion

We have implemented video retrieval and summarization for a home with a large number of sensors, by analyzing both content and context data. While the inclusion of multiple sensory modalities and the human centered aspect of the system made the research fairly broad, it was possible to make a number of novel contributions.

Hierarchical clustering of floor sensor data followed by video handover enabled the creation of personalized video clips using a large number of cameras. It was possible to dub this video with reasonably good quality, using audio handover. An adaptive algorithm enabled retrieval of more than 80% of the key frames required for a complete summary of the video.

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 achieve generally accurate sound source localization despite the absence of microphone arrays or a beam-forming setup. The accuracy of audio classification using only time domain features is above 83%, suggesting that high accuracy of classification is possible at the expense of further analysis using features from multiple domains.

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 user interface based on hierarchical media segmentation and Interactive retrieval facilitated 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 helped the user to retrieve required media.

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.

10.2 Future Work

The accuracy of the existing algorithm for footstep segmentation can be improved for more efficient retrieval of personalized video. Adding person recognition capability to this stage will enable the creation of *personal diaries* for residents.

Use of frequency and Cepstral domain features for similarity matching is a prospective approach for improving the performance of sound source localization. Applying Independent Component Analysis (ICA) before Audio classification to separate multiple simultaneous sound sources within the same region is an interesting research direction.

Further analysis of sensor data, especially image data, can be used for detection and recognition of higher-level actions and events, such as conversations, mealtimes, etc., thereby enhancing the functionality of the system. Face detection in retrieved images and video can provide additional information for searching within the data.

Novel techniques for user interaction and visualization of results can be designed, to achieve more effective and efficient retrieval. Continuous evaluation of the system

based on user studies and feedback is vital, as usability is one of the most important criteria for an effective retrieval system for the home.

References

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

[2] Pentland, A. “Smart Rooms,” Scientific American, pp.54-62, 1996.

[3] Harper, R., “Inside the Smart Home: Ideas, Possibilities, and Methods”, Springer-Verlag UK, July 2003.

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

[5] Noguchi, K.; Somwong, P.; Matsubara, T.; Nakauchi, Y., "Human intention detection and activity support system for ubiquitous autonomy," C*omputational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on* , vol.2, no.pp. 906- 911 vol.2, 16-20 July 2003.

[6] Brumitt, B., “Easy Living: Technologies for Intelligent Environments,” Proc. of International Symposium on Handheld and Ubiquitous Computing, 2000.

[7] Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S. “Multi-Camera Multi-Person Tracking for Easy Living,” Proc. of 3rd IEEE International Workshop on Visual Surveillance, pp.3-10, 2000.

[8] Lee, J., Ando, N., Hashimoto, H., “Design Policy for Intelligent Space,” Proc. of IEEE SMC’99, 1999.

[9] Juels, A., "RFID security and privacy: a research survey," *Selected Areas in Communications, IEEE Journal on* , vol.24, no.2pp. 381- 394, Feb. 2006

[10] Fishkin, K. P., Wang, M., Borriello, G., “A ubiquitous system for medication monitoring,” Intel Res., Seattle, Tech. Memo IRS-TR-03-011, Oct., 25 2004.

[11] Lee, S., Mase, K.,. “Activity and Location Recognition Using Wearable Sensors”, Pervasive Computing, July-September, pp.10-18, September, 2002.

[12] Petrushin, V. A., Gang W., Omer S., Damian R., Gershman, V., "Multiple-Sensor Indoor Surveillance System," *crv*, p. 40, The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), 2006.

[13] Lee, D., Hull, J. J., Erol, B., “Meeting Video Retrieval using Dynamic HMM Model Similarity”, IEEE ICME Conference, 2005.

[14] Jaimes, A., Nagamine, T., Liu, J., Omura, K., Sebe, N., “Affective Meeting Video Analysis”, IEEE International Conference on Multimedia and Expo (ICME’05), Amsterdam, The Netherlands, July 2005.

[15] Lijun T., Kender, J.R., "Semantic Indexing for Instructional Video Via Combination of Handwriting Recognition and Information Retrieval," *Multimedia*

and Expo, 2005. ICME 2005. IEEE International Conference on , vol., no.pp. 920- 923, 6-8 July 2005

[16] Bibiloni, A., & Galli, R. (2003) Content Based Retrieval Video System for Educational Purposes. *Proceedings of Eurographics Workshop on Multimedia "Multimedia on the Net, EGMM 1996*,

[17] Waibel, A., "CHIL – Computers in the Human Interaction Loop", In proceedings of LEARNTEC 2005.

[18] Cucchiara, R. 2005. Multimedia surveillance systems. In Proceedings of the Third ACM international Workshop on Video Surveillance &Amp; Sensor Networks (Hilton, Singapore, November 11 - 11, 2005). VSSN '05. ACM Press, New York, NY, 3-10.

[19] Gandhi, T., Trivedi, M. M., "Calibration of a reconfigurable array of omnidirectional cameras using a moving person, Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks, October 15-15, 2004, New York, NY, USA

[20] Lam, K., Chiu, K. H. C., Adaptive visual object surveillance with continuously moving panning camera, Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks, October 15-15, 2004, New York, NY, USA

[21] Valera, M., Velastin, S.A., Intelligent distributed surveillance systems: a review vision. In Image and Signal Processing, IEE Proceedings, volume 152, pages 192--204, April 2005.

[22] Y.A. Ivanov and A.F. Bobick, Recognition of Multi-agent Interactions in Video Surveillance," IEEE Proceedings of the International Conference on Computer Vision, Kerkyra, Greece, Vol. 1, pp. 169-176, September 20-27, 1999.

[23] J. Davis and A. Tyagi, "Minimal-Latency Human Action Recognition using Reliable-Inference", *Image and Vision Computing*, Vol 24, 2006, pp. 455-472.

[24] Collins, R.T., Lipton, A.J., Fujiyoshi, Kanade, T., "Algorithms for cooperative multisensor surveillance", In Proc. of the IEEE, volume 89, pp. 1456-1477, Oct. 2001.

[25] Brooks, R.R., Ramanathan, P., Sayeed, A.M., "Distributed target classification and tracking in sensor networks. In Proc. of the IEEE", volume 91, pp. 1163-1171, 2003.

[26] Arampatzis, Th.; Lygeros, J.; Manesis, S., "A Survey of Applications of Wireless Sensors and Wireless Sensor Networks," *Intelligent Control, 2005. Proceedings of the 2005 IEEE International Symposium on, Mediterranean Conference on Control and Automation* , pp. 719- 724, 2005

[27] Khalifa, Y.; Okoene, E., "A distributed agent-based surveillance system," *Signals, Circuits and Systems, 2005. ISSCS 2005. International Symposium on*, vol.2, no.pp. 713- 716 Vol. 2, 14-15 July 2005

[28] Girgensohn, F., Shipman, S., Dunnigan, A., Turner, T., Wilcox, L., "Support for Effective Use of Multiple Video Streams in Security" , In Proceedings of the

Fourth ACM international Workshop on Video Surveillance & Sensor Networks , VSSN '06. ACM Press, New York, NY, pp.19-26.

[29] Wang, Y., Liu, Z., Huang, J., “Multimedia Content Analysis Using Both Audio and Visual Cues”, IEEE Signal Processing Magazine, November 2000. 12-36

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

[31] Cai, R., Lu, L., Hanjalic, A., “Unsupervised Content Discovery in Composite Audio”, Proceedings of ACM Multimedia 2005, Singapore. (2005) Page 628-637

[32] Pye, D., Hollinghurst, N., Mills, T. and Wood, K., "Audio-visual Segmentation for Content-based Retrieval", The International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia, December 1998

[33] Sung I. Yong , Won S. Lee, Content-based retrieval of video data with flexibly managed attributes, Knowledge and Information Systems, v.4 n.4, p.507-519, October 2002

[34] Kim, S., Hwang, D. S., Kim, J., Seo, Y., “An Effective News Anchorperson Shot Detection Method Based on Adaptive Audio/Visual Model Generation”, Proceedings of the Fourth International Conference in Image and Video Retrieval (CIVR 2005), Singapore. (2005) 276-285

[35] Brown, M.G., Foote, J.T., Jones, G.J.F., Sparck Jones, K., and Young, S.J., 1995, "Automatic Content-based Retrieval of Broadcast News," Proceedings of ACM Multimedia. San Francisco: ACM, pp. 35 - 43.

[36] Chien Yong Low , Qi Tian , Hongjiang Zhang, An automatic news video parsing, indexing and browsing system, Proceedings of the fourth ACM international conference on Multimedia, p.425-426, November 18-22, 1996

[37] Duan, L., Xu, M., Tian, Q., Xu, C., Jin, J.S. “A unified framework for semantic shot classification in sports video”, IEEE Transactions on Multimedia, Volume: 7, Issue: 6, Dec. 2005. pp.1066-1083

[38] Tjondronegoro, Dian and Chen, Yi-Ping Phoebe and Pham, Binh (2002) *A Framework for Customizable Sport Video Management and Retrieval*, in Za¨ıane, Osmar and Simoff, Simeon and Djeraba, Chabane, Eds. *Mining Multimedia and Complex Data: KDD Workshop MDM/KDD 2002, PAKDD Workshop KDMCD 2002 Revised Papers*, pages pp. 248-265. Springer.

[39] Huayong, L. and Hui, Z. 2005. A Content-Based Broadcasted Sports Video Retrieval System Using Multiple Modalities: SportBR. In *Proceedings of the Fifth international Conference on Computer and information Technology* (September 21 - 23, 2005). CIT. IEEE Computer Society, Washington, DC, 652-656.

[40] Petkovic, M., Jonker, W., “Cobra: A Content-Based Video Retrieval System”, Advances in Database Technology - EDBT 2002 : 8th International Conference on Extending Database Technology, Prague, Czech Republic, March 25-27, 2002. Proceedings, pp. 736-738.

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

[42] 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.

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

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

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

[46] Bush, V., :As We May Think", The Atlantic Monthly, 176(1), July 1945, pp.101-108.

[47] S. Mann, "Humanistic Intelligence: `WearComp' as a new framework and application for intelligent signal processing", Proceedings of the IEEE, Vol. 86, No. 11, November, 1998

[48] Jennifer Healey , Rosalind W. Picard, StartleCam: A Cybernetic Wearable Camera, Proceedings of the 2nd IEEE International Symposium on Wearable Computers, p.42, October 19-20, 1998

[49] Wood, K., Fleck, R. & Williams, L (2004) Playing with SenseCam. Proceedings of Playing with Sensors: Exploring the boundaries of sensing for playful ubiquitous computing (W3), UbiComp 2004

[50] Gemmell, J., Bell, G., Lueder, R., Drucker, S., Wong, C., "MyLifeBits: fulfilling the Memex Vision", in Proceedings of the tenth ACM international conference on Multimedia, ACM Press: Juan-les-Pins, France. pp. 235-238.

[51] Michael Beigl, Albert Krohn, Tobias Zimmer and Christian Decker: Typical Sensors needed in Ubiquitous and Pervasive Computing First International Workshop on Networked Sensing Systems (INSS) 2004, Tokyo, Japan, June 22-23. 2004, pp 153-158

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

[53] Gemmell, J., Bell, G., Lueder, R., "MyLifeBits: a personal database for everything", Communications of the ACM, vol. 49, Issue 1 (Jan 2006), pp. 88-95.

[54] Ubiquitous Home: http://www.nict.go.jp/jt/a135/eng/research/ubiquitous_home.html, National Institute of Information and Communication Technology, Japan.

[55] Smeaton, A.F., McHugh, M., "Towards Event Detection in an Audio-Based Sensor Network", Proceedings of the Third ACM Workshop on Video Surveillance and Sensor Networks (VSSN'05), Singapore. (2005). pp.87-94

[56] Rui, Y., Florencio, D., “New direct approaches to robust sound source localization”, IEEE International Conference on Multimedia & Expo (ICME 2003), Baltimore, USA. (2003)

[57] Chen, J. F., Shue, L., Sun, H. W., Phua, K. S. “An Adaptive Microphone Array with Local Acoustic Sensitivity”, IEEE International Conference on Multimedia & Expo (ICME 2005), Amsterdam, The Netherlands. (2005)

[58] Hoshuyama, A. Sugiyama, A. Hirano, T., “A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters”, IEEE Trans. Signal Processing, vol. 47, no. 10, pp. 2677--2684, Oct. 1999.

[59] Vacher, M., Istrate, D., Besacier, L., Castelli, E., Serignat, J. F., “Smart audio sensor for telemedicine”, Smart Objects Conference 2003, Grenoble, France. (2003)

[60] Bian, X., Abowd, G. D., Rehg, J. M., “Using Sound Source Localization to Monitor and Infer Activities in the Home”, GVU Technical Report; GIT-GVU-04-20, Georgia Institute of Technology. (2004)

[61] Liu, M., Wan, C., “A Study on Content-Based Classification and Retrieval of Audio Database”, Proceedings of the 2001 International Database Engineering and Applications Symposium, Grenoble, France. (2001) pp. 339-345

[62] Vazirgiannis, M., Sellis, T., "*Event and Action Representation and Composition for Multimedia Application Scenario Modelling*", Proceedings of the European Workshop on Interactive Distributed Multimedia Systems and Services (IDMS'96), March 1996.

[63] Snoek, G.M., Worring, M., “Multimedia Event-Based Video Indexing using Time Intervals”, IEEE Transactions on Multimedia, 7(4) August 2005. pp.638-647,

[64] Kraft, F., Malkin, R., Schaaf, T., Waibel, A., “Temporal ICA for Classification of Acoustic Events in a Kitchen Environment,” in INTERSPEECH, Lisbon, Portugal, 2005.

[65] de Silva, G.C., “Tracking and Indexing of Human Actions in Video Image Sequences”, M. Eng. Thesis, Dept. of Electrical and Computer Engineering, the National University of Singapore, 2003.

[66] Wang, J. R. and Parameswaran, N. 2004. Survey of sports video analysis: research issues and applications. In *Proceedings of the Pan-Sydney Area Workshop on Visual information Processing*

[67] Piccardi, M., Hintz, T., He, S., Huang, M. L., Feng, D. D., Eds. ACM International Conference Proceeding Series, vol. 100. Australian Computer Society, Darlinghurst, Australia, pp.87-90.

[68] Witten, I. H., Frank, E., (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[69] Viola, P., Jones, M., *Robust Real-time Object Detection*, International Journal of Computer Vision, 2002. pp.1026-1033

[70] Anavi, S., “Information Processing in a Ubiquitous Home Using Image Analysis”, Masters Thesis, Signal Processing Institute, School of Engineering Swiss Federal Institute of Technology, Lausanne, February 2006.

[71] Zhao, R., and Grosky, W. I., “Negotiating The Semantic Gap: From Feature Maps to Semantic Landscapes”, Pattern Recognition, Volume 35, Number 3 (March 2002), pp. 51-58.

[72] Sommerville, I., “Software Engineering: 5th Edition”, Addison-Wesley, 1993.

[73] Pingali, G., “eChronicles: Scope, Opportunities, and Challenges”, Proceedings of IEEE eChronicle Workshop, 2006, pp.1-6.

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

Appendix A

Material used for Evaluation of Key Frame Extraction

Annexed starting from the following page are the instruction sheet, handouts and answer sheets used for the experiment for evaluating the algorithms for key frame extraction (Section 5.6). The key frames sets and the answer sheets for only one video clip are included, as an example.

Instruction Sheet

Overview

The objective of this experiment is to evaluate different methods of extracting *key frames* from a video clip. Key frames, in our context, are frames that are selected such that they present a **summary** of the events took place in the video clip. The motivation for creating a key frame set is saving time taken to view an entire video clip to find out what happened.

Please follow these instructions while taking part in the experiment. If there is anything that is not clear enough, please ask.

Instructions

- Before you start, please fill in your name and start time of the experiment, in the spaces provided in the answer sheet.
- You have to repeat the following steps on four video clips (sequences). You can take a break after each sequence.

1. Use the *Image Sequence Viewer* to watch the sequence. This is the original video, which is not summarized yet. Try using the buttons for playing, pausing and moving back and forth along the sequence. The durations of the sequences are:
    - Sequence 1 - 36 seconds
    - Sequence 2 - 1 minute
    - Sequence 3 - 3 minutes
    - Sequence 4 - 5 minutes
2. Select key frames for the sequence so that they represent a summary of the behavior of the person shown in the video clip. There is no restriction on the number of frames you can choose. There is no time limit for selecting key frames. Write the timestamps of the frames, in the sheet provided.

    Notes:

    - It is not necessary to write down the date component of the timestamp
        - Example: **2004/09/03 13:2:45:13** can be written down as **13:2:45:13**
    - It is not necessary to write the timestamps in any particular order. Feel free to revise and add more key frames. You may strike out the entries you made (Example: ~~13:12:24:5~~) if you feel that they are not necessary.

3. Now, observe each set of printed key frames and fill in the corresponding column in the table in Section A. There are seven sets of key frames for each sequence. You can go back to view the sequence as many times as necessary.
4. Proceed to Section C only after completing Section B.

- At the end of the experiment, please record the time in the space provided.

Experiment for Evaluation of Key Frame Extraction

Name : _____

Date : _____

Start time : _____

End time : _____

Test Sequence 1

Please use the numbers below the images to specify key frames.

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

Sequence 1

Section A

Key frames selected for this sequence

Timestamp	Timestamp

Section B

Evaluation of individual key frame sets

	Key frame set						
	A	B	C	D	E	F	G
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							

Section C

2. Out of the 7 frame sets, which one do you think best key frame set for this sequence?

3. For the selected frame set:

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

   (b) What are the ways that it can be improved?

4. Additional comments for this sequence and key frame sets, if any:

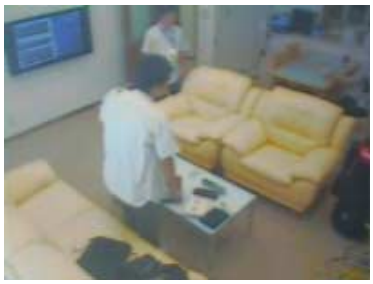
Sequence 1

Frame set A

2004/9/3 13:41:2:28

2004/9/3 13:41:4:28

2004/9/3 13:41:5:28



2004/9/3 13:41:15:11

2004/9/3 13:41:29:17

2004/9/3 13:41:30:23

2004/9/3 13:41:33:23

Sequence 1

Frame set B

2004/9/3 13:41:1:22

2004/9/3 13:41:3:10

2004/9/3 13:41:5:4

2004/9/3 13:41:25:17

2004/9/3 13:41:30:23

2004/9/3 13:41:31:17

Sequence 1

Frame set C

2004/9/3 13:41:2:28

2004/9/3 13:41:4:28

2004/9/3 13:41:5:28

2004/9/3 13:41:31:29

2004/9/3 13:41:32:29

2004/9/3 13:41:38:29

Sequence 1

Frame set D

2004/9/3 13:41:1:16

2004/9/3 13:41:3:16

2004/9/3 13:41:5:16

2004/9/3 13:41:20:17

2004/9/3 13:41:30:17

2004/9/3 13:41:31:17

Sequence 1

Frame set E

2004/9/3 13:41:1:16

2004/9/3 13:41:3:16

2004/9/3 13:41:5:16

2004/9/3 13:41:30:17

2004/9/3 13:41:31:17

Sequence 1

Frame set F

2004/9/3 13:41:2:28

2004/9/3 13:41:17:17

2004/9/3 13:41:32:11

2004/9/3 13:41:38:29

Sequence 1

Frame set G

2004/9/3 13:41:2:28

2004/9/3 13:41:32:11

2004/9/3 13:41:38:29

Appendix B

Simplified Mathematical Model for Sound Source Localization

This appendix attempts to justify the scaled template matching using the energy distribution templates, using a simplified approach.

Let us assume that the sound energy received by the microphones within a given region of a closed environment is a linear combination of sound energy generated in each region, and no noise is present. The sound energy received within region m, r_m can be stated as

$$r_m = a_{m1}s_1 + a_{m2}s_2 + \ldots + s_m \ldots + a_{(n-1)}s_{(n-1)} + a_{mn}s_n$$

where n is the number of sources in the closed environment and S_i ($i = 1..n$) is the sound energy released by the i^{th} sound source.

The ubiquitous home is partitioned into 7 regions where sounds can be generated. Sounds are captured by 25 microphones in 6 of those regions. For simplicity, we assume that we have only one *receiver* in each of these regions. The equations for received sound energy by the set of receivers can now be written in matrix form as

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \end{bmatrix} = \begin{bmatrix} 1 & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} \\ a_{21} & 1 & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} \\ a_{31} & a_{32} & 1 & a_{34} & a_{35} & a_{36} & a_{37} \\ a_{41} & a_{42} & a_{43} & 1 & a_{45} & a_{46} & a_{47} \\ a_{51} & a_{52} & a_{53} & a_{54} & 1 & a_{56} & a_{57} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & 1 & a_{67} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \end{bmatrix}$$

$\Rightarrow \mathbf{r} = \mathbf{As}$

The matrix **A** represents the proportion of sound energy that propagates from one region to the other, in a normalized form. If matrix **A** can be obtained, the problem of sound source localization can be reduced to solving this matrix equation for **s** given **r** and **A**.

The energy distribution templates for the regions of ubiquitous home represent the propagation of sound energy to individual microphones of each region. Hence, each of the energy distribution templates serve as a detailed representation of each column of **A**.

However, it should be noted that the system cannot be solved by matrix inversion, since the number of elements in **r** is smaller than that of **s**. The scaled energy distribution template matching technique is equivalent to solving the set of linear equations for **s**.

Appendix C

Material used for the User Study

The annexed answer sheet was provided to the participants of the user study described in Chapter 9.

Answer Sheet

Name: .....

Part A: Requirements study

1. Suppose it is possible to retrieve any event that happened anytime during your stay in ubiquitous home. What are the things you would like to see from that stay? Please check any number of items as you like, and add your own choices.

   - ❑ Things that I did
   - ❑ Things that the other family members did
   - ❑ Things we did together
   - ❑ How my child was playing when she was alone
   - ❑ Things that I have forgotten
   - ❑ A summary of what I did each day
   - ❑ .....

     .....
   - ❑ .....

     .....
   - ❑ .....

     .....
   - ❑ .....

     .....

2. Supposing it is possible to have the same facility at your home, and only your family has access to the data:

(a) What would you like to use it for?

(b) How would you like to record the data?

- ❑ Record everything from daily life so that I can find out anything if I forget
- ❑ Record only the special events such as (please specify):

(c) Which parts of the house will you record everything?

- ❑ Everywhere
- ❑ The following places (please specify):

……

……

(d) Which parts of the house will you refrain from recoding?

……

……

(e) Do you have any other preferences, such as times of day, about recording?

……

……

Part B: Usability Ratings

Please answer the following questions after the demo and the hands-on session. Circle the number corresponding to your ranking for each criterion

Learning to use the system

1 Difficult	2	3	4	5	6	7 Easy

Overall impression

1 Difficult to use	2	3	4	5	6	7 Easy to use

1 Useless	2	3	4	5	6	7 Very useful

Part C: Descriptive Feedback

Please be descriptive as and when necessary when answering the following questions:

1. How much did you remember from what you could see in the video and key frames?

2. Was it possible to see interesting things that you did not see/know before?

3. Out of what you saw, which parts of video would you like to keep with you?

4. State what you like about this software.

5. State what you don't like about it.

6. For what kind of things will this software be useful to you?