9

Recognition of Real-World Activities from Environmental Sound Cues to Create Life-Log

Mostafa Al Masum Shaikh, Keikichi Hirose and Mitsuru Ishizuka
Dept. of Information and Communication Engineering,
The University of Tokyo, Tokyo,
Japan

1. Introduction

There are several studies that collect and store life-log for personal memory. This chapter explains about a system that can create someone's life-log in an inexpensive way to share daily life events with family, friends or care-givers through simple text messaging with a notion to remote monitoring of someone's wellbeing. In the developed world where people are usually busier than ever, ambient communications through mobile media or the Internet based communication can provide rich social connections to their loving ones ubiquitously whom they care about by sharing awareness information in a passive way. For users who wish to have a persistent existence through ambient communication – to let someone else to know about their daily activity – new technology is needed. Research that aims to simulate virtual living or logging daily events, while challenging and promising, is currently rare. Only very recently the detection of real-world activities has been attempted by processing multiple sensors data along with inference logic for real-world activities. Detecting or inferring human activity using such simple sensor data is often inaccurate, insufficient and expensive. Therefore, this chapter discusses a technology, an inexpensive alternative to other sensors (e.g., accelerometers, proximity sensors etc.) based approaches, to infer human activity from environmental sound cues and common-sense knowledgebase of everyday objects and concepts. A system prototype to log daily events to infer activities in 'as you go' manner from environmental sound cues is explained with a few case studies. The input of the system is the patterns of sounds that are usually produced from activities (e.g., toilet flushing), occurring environmentally (e.g., road sounds) or due to interaction with the objects (e.g., cooking utensils clattering). A robust signal processing processes the input sound signal and Hidden Markov Model (HMM) classifiers are developed to detect pre-determined sound contexts. Based on the detected sounds and along with the common-sense knowledge regarding human activity, object interaction, ontology of human life (e.g., living pattern of a single old man, or an old couple) and temporal information (e.g., morning, noon etc.) inference engine is employed to detect the activity and the surrounding environment of the person. Preliminary results are encouraging with the accuracy rate for outdoor and indoor related sound categories for activities being above 67% and 61% respectively.

2. Environmental sound cues and life-log

Although speech is the most informative acoustic event, but environmental sounds are also useful to process because those can provide useful information regarding context of the environment. In a given environment human-activity can be reflected by a variety of acoustic events, either produced naturally or by the human body or by the objects manipulated or interacted. For example: Jingling sound of cooking utensils (like cooking pan, spoon, knife etc.) may lead to infer someone's cooking activity, likewise vehicle passing sound may lead to infer that someone is on the road, etc. Many sources of information for sensing the environment as well as activity are available (Chen et al., 2005; Philipose et al., 2004; Temko and Nadeu, 2005). In this chapter, we consider two objectives namely, sound-based context awareness, where the decision is based merely on the available acoustic information at the surrounding environment of the user and automatic life-logging, where the detected sound context infers an activity to be logged along with temporal information. Acoustic Event Detection (AED) is a recent sub-area of computational auditory scene analysis (Wang and Brown, 2006) that deals with the first objective. AED processes acoustic signals and converts those into symbolic descriptions corresponding to a listener's perception of the different sound events that are present in the signals and their sources. Life-log is a chronological list of activities performed by the user with respect to time. Such a list might indicate the user's well-being or abnormality according to the consideration of the person's self assessment or by someone else who cares about the person (e.g., relatives or care-givers). Therefore we apply the concept of AED to perform automatic generation of life-log. This life-log can be transmitted autonomously as a simple text message to someone else with the notion of ambient communication.

Life logs include people's activities performed in specific locations at a specific time and it can be collected from various sources. We envisage that with the proliferation of computing power of hand held devices (HHD), availability of the Internet connectivity and improvements in communication technologies ambient communication will find a universal place at our daily life and allow us to realize virtual living through ambient social communication. Let's consider the following scenario of a globalized family.

Scenario 1: Rahman family (Mr. and Mrs. Rahman) lives in Khulna, one of the metropolitan cities of Bangladesh. They have three sons living overseas, one in Texas, another in Ottawa and the youngest one in Bonn of Germany. Both Mr. and Mrs. Rahman are now at their age of over 50 and Mr. Rahman had a massive heart operation last year. Mrs. Rahman is also ailing from several sicknesses like diabetics, high blood pressure, etc. The three sons are always worried regarding the well being of their parents and consequently they often talk to their parents over the phones to know their whereabouts. Though calling to Khulna, Bangladesh from USA, Canada, and Germany is relatively cheaper now-a-days than before, but having a phone conversation with their parents is not always possible due to various reasons, for example, due to inconvenience in time differences (e.g., when it is 10 am in Khulna it is 11:00 pm in Texas, 12:00 am in Ottawa and 6:00 am in Bonn) that is, when the sons have convenient time to call, their parents are usually sleeping or resting. But they are often worried to know at least how their parents are doing everyday. Therefore, let's imagine that Rahman family has internet connectivity at their home and installed an inexpensive system capable of doing the followings. The system makes automatic life-logging of daily activities by detecting and recognizing sound cues from their surrounding environments and sends email message(s) to their sons reporting their daily life-sketch. In

this case, an example email message containing life log for a particular day, as follows, might be very relieving to the sons. *"Your parents woke up at 7:30 am today and they had breakfast around 8:15 in the morning. They watched TV several times in the day. Went out of home for two times and walked in the roads and parks. They took lunch and dinner at around 2 PM and 8 PM. Your mother went to toilet for 5 times and father went to toilet 6 times in a day. They had communicated with each other or other people by talking. It seems they are doing fine."*

In this chapter, we describe a listening test made to facilitate the direct comparison of the system's performance to that of human subjects. A forced choice test with identical test samples and reference classes for the subjects and the system is used. The second main concern in this chapter is to evaluate how acceptable the automatic generation of life-log is. Since we are dealing with a highly varying acoustic material where practically any imaginable sounds can occur, we have limited our scope in terms of location and the activities to recognize at a particular location. It is most likely that the acoustic models we are using are not able to sufficiently model the observation statistics. Therefore, we propose using discriminative training instead of conventional maximum likelihood training.

The chapter is organized as follows: Section 3 reviews the background studies related to this research. Our approach, in terms of system architecture and description of the system components is explained in Section 4. Section 5 discusses about several development challenges and our response towards those. Section 6 explains the experimental setup, the results obtained by the detection and classification system as well as user evaluations. Conclusions are presented in Section 7.

3. Background

A number of researchers have investigated to infer activities of daily living (ADL). Mihailidis et al. (2001) have successfully used cameras and a bracelet to infer hand washing. Wan (1999) used radio-frequency-identification (RFID) tags functionally as contact switches to infer when users took medication. The system discussed in (Barger et al., 2002) used contact switches, temperature switches, and pressure sensors to infer meal preparation. Tran et al. (2001) used cameras to infer meal preparation. Glascock and Kutzik (2000) used motion and contact sensors, combined with a custom-built medication pad, to get rough inference on meal preparation, toileting, taking medication, and up-and-around transference. A custom wearable computer with accelerometers, temperature sensors, and conductivity sensors to infer activity level is used in (Korhonen et al., 2003). Mozer (1998) used 13 sensors to infer home energy use, focusing on the heating-use activity. Motion detectors to infer rough location were used in (Campo and Chan, 2002). Several sensors like motion sensors, pressure pads, door latch sensors, and toilet flush sensors to infer behavior are reported in the system described in (Guralnik and Haigh, 2002). Chen et al. (2005) have described monitoring bathroom activities based on sound. The system (Philipose et al., 2004) utilized RFID tags to detect objects and thereby inference of activities is done from the interaction with the detected objects. The research on MIT's *house_n* project (MIT house_n, 2008) places a single type of object-based adhesive sensor in structurally unmodified homes and sensor readings are later analyzed for various applications—kitchen design, context sampling, and potentially ADL monitoring. All of these systems have a commonality that they perform high-level inference from low-level by coarse sensor data reporting and analyses. Some have added special pieces of hardware to help performance improvement, but progress toward accurate ADL detection has nevertheless been slow. Only a few researchers have reported

the results of any preliminary user testing (Mihailidis et al., 2001; Glascock and Kutzik, 2000; Campo and Chan, 2002; Guralnik and Haigh, 2002). The level of inference using sensors has often been limited, for example, reporting only that a person entered the living room and spent time there. Moreover, as an example, research aiming to detect hand washing or tooth brushing have had nearly no synergy, each using its own set of idiosyncratic sensors and algorithms on those sensors. Furthermore a home deployment kit designed to support all these ADLs would be a mass of incompatible and non-communicative widgets. Our approach instead focuses on a general inference engine and infers activities from the sound cues that are likely to be produced either naturally or from the interactions with objects. Thus we can use our system for a broader range of ADLs.

The idea of a "life-log" or a personal digital archive is a notion that can be traced back at least 60 years (Bush, 1945). Since then a variety of modern projects have spawned such as the *Remembrance Agent* (Rhodes and Starner, 1996), *the Familiar* (Clarkson and Pentland, 1999; Clarkson et al., 2001), *myLifeBits* (Gemmell et al, 2002), *Memories for Life* (Fitzgibbon and Reiter, 2003) and *What Was I Thinking* (Vemuri and Bender, 2004). In (Blum et al., 2006) the authors evaluate the user's context in real time and then use variables like current location, activity, and social interaction to predict moments of interest. Audio and video recordings using a wearable device can then be triggered specifically at those times, resulting in more interest per recording. Some previous examples of this approach are the Familiar and *iSensed* systems (Clarkson and Pentland, 1999; Clarkson et al., 2001; Blum et al., 2006) which structure multimedia on the fly; the *eyeBlog* system (Dickie et al., 2004) which records video each time eye contact is established; and the *SenseCam* (Gemmell et al, 2004), which records images and sound whenever there's a significant change in the user's environment or the user's movement. Life log includes people's experiences which are collected from various sensors and stored in mass storage device. It is used to support user's memory and satisfy user's needs for personal information. If he wants to inform other people of his experience, he can easily share his experience with them by means of providing his life log.

To collect life log (e.g., GPS based location, SMS, call, charging, MP3, photos taken, images viewed, and weather information, etc) smart phones (e.g., iPhone 3G) are usually used. Smart phone is a mobile device that includes color LCD screen, mass storage, large memory, and communicative function by using Wi-Fi, Bluetooth, and infrared. It also has a variety of softwares like scheduler, address book, media player, and e-book. Raento et al., (2005) developed a framework for collecting contexts from smart phone which collects GSM Cell ID, Bluetooth, GPS data, phone data, SMS data, and media information that are transmitted to the server. The contexts could be provided for other contents as additional information. Panu *et al.* collect log data from mobile devices, and extracts features by pre-processing the log data (Panu et al., 2003). The mobile device uses GPS log, microphone, temperature, moisture, and light sensor. *MyLifeBits* Project is one of the implementations of personal record database system (Gemmell et al, 2002). Personal information is collected by PC, *SenseCam* and so on, and stored in database server with relationships among personal information. However, user faces difficulties to explore and search contents because of large amount of personal data. KeyGraph-based mobile contents management system was suggested to manage user's information in mobile device, which extracted important information using KeyGraph algorithm and provided searching or exploring contents (Kim et al., 2007). The problem of the system is using only log data. If analysis and inference of the data was added to the system, it would give better performance.

Our work differs from others in three key ways. First, we utilize environmental sounds cues to infer the interactions with objects or environment instead of sensor or camera data. Thus we can identify a large set of objects like spoons, toothbrushes, plates etc. Second, due to simple use of portable microphone to capture environmental sound we can also infer outdoor environments like on the road, in a park, in a train station etc., that previous research was limited to perform. Thirdly, our model is easy to incorporate new a set of activities for further needs by just adding more appropriately annotated sound clips and re-training the HMM based recognizer.

4. Our approach

Our approach to log daily events employs the technique to detect activities of daily living (e.g., laughing, talking, traveling, cooking, sleeping, etc.) and situational aspects of the person (e.g., inside a train, at a park, at home, at school, etc.). The system infers human-activity from environmental and object-interaction related sound cues as well as common-sense knowledge. Initially, with a view to creating a sound corpus of environmental sounds, 114 types of acoustic sounds are collected that are usually produced during object interaction (e.g., cooking pan jingling sound while cooking) or by the environment itself (e.g., bus/car passing sound while on a road) or by a deliberate action of a person (e.g., laughing, speaking). This sound corpus serves the purpose to train the system to infer a person's activity and one's surroundings with the help of common-sense knowledge. For example, for a sound-clip recorded from the environment at a particular time if the system identifies cooking pan's jingling and chopping-board's sound as consecutive cues and the system's local time indicates evening then from common-sense database the system infers this activity as 'cooking' for the input.

4.1 System architecture

The top-level pipelined architecture of the system is portrayed by Figure 1. Due to the rapid development and popularity of Hand Held Devices (HHD), HHD (e.g., portable computer or smart phone) has been considered as an implementation medium to deploy this application as an "always on" type system. The application pays heed to the environment to capture environmental sound continually with due intervals (in this case 10 seconds) and thereafter that recorded sound-clips are be processed. According to Figure 1, a sound-clip resulted from the listening of the environment, is passed through a robust signal processing and sound cues are detected by the trained HMM classifiers. HMM classifiers are trained using the collected sound corpus. Based on the detected sound cues and common-sense knowledge regarding human activity, object interaction, ontology of human life (e.g., daily life of a student, or a salary man etc.) and temporal information (e.g., morning, noon etc.) are applied to infer the activity and the surrounding environment of the person. This information is then stored in the log of activities as English sentences.

4.2 Description of system components

In this section the system components are described briefly.

4.2.1 Sound corpus

The patterns of sounds arising from activities occurring naturally or due to interaction with the objects are obviously a function of a many environmental variables like size and layout

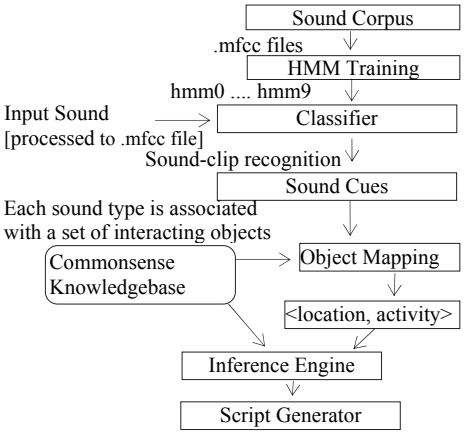


Fig. 1. The System's Architecture

or the indoor environment, material of the floors and walls, type of objects (e.g., electrical or mechanical) and persistent ambient noise present in the environment etc. In is essential to install this system to the same culture and environment from where sound samples are acquired and proper training of the system is made. It is analogous to the practice adopted for speech recognition whereby the system is individually trained on each user for speaker dependent recognition because such environmental sounds may vary in different cultures and places. Therefore the sample sounds we have collected are from the different places of Tokyo city and Tokyo University. For clear audio-temporal delineation during system training, the sound capture for each activity of interest was carried out separately. A number of male and female subjects were used to collect the sounds of interest; each subject would typically go into the particular situation as depicted in Table I with the sound recording device and the generated sounds are recorded. We used the digital sound recorder of SANYO (model number: ICR-PS380RM) and signals were recorded as Stereo, 44.1 KHz, .wav formatted files. It is important to note that in the generation of these sounds, associated 'background' sounds such as the ambient noise, rubbing of feet, friction with cloths, undressing, application of soap, etc., are being simultaneously recorded. The

Location	Activities
Living Room	Listening Music, Watching TV, Talking, Sitting Idle, Cleaning (e.g., by vacuum-cleaning)
Work Place	Sitting idle, Working with PC, Drinking
Kitchen	Cleaning, Drinking, Eating, Cooking
Toilet	Washing, Urinating, Flushing out
Gym	Exercising
Train Station	Waiting for Train
Inside Train	Traveling by Train
Public Place	Shopping, Traveling on Road
On the Road	Traveling on Road

Table 1. List of locations and activities of our interest

variability in the captured sounds of the each activity provides realistic input for system training, and increases the robustness and predictive power of the resultant classifier. Some sounds (e.g., water falling, vacuum cleaning machine sounds etc.) are generally loud and fairly consistent. There are samples that needed to sufficiently train the classification model due to a high degree of variability even for the same individual. For example, hands washing, drinking, eating, typing related sounds exhibited a high degree of variability. This required us to collect many more samples for such kind of activities related sounds to capture the diversity of the sounds. According to the location and activities of our interest mentioned in Table I, we have collected 114 types of sounds. Each of the sound types has 15 samples of varying length from 10 second to 25 seconds.

The typical waveforms and energy spectrum of the water related sounds of three different actions are shown in Fig. 2. As can be seen, the water splashing on hand or face (frame 1), toilet flushing (frame 2), shower water (frame 3), the sounds as depicted in the waveforms and spectrum have distinct patterns, different durations, energy spectrum and amplitudes.

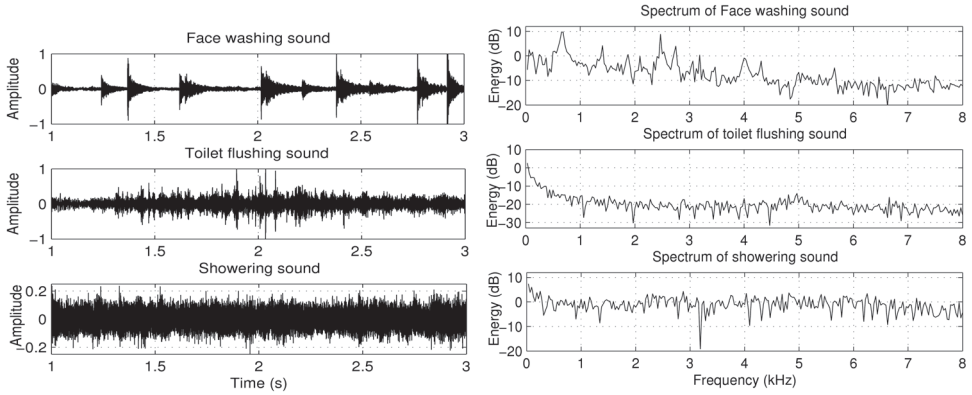


Fig. 2. Waveforms of different water related sound types

4.2.2 HMM training

An accurate and robust sound classifier is critical to the overall performance of the system. There are however many classifier approaches in the literature, e.g., those based on Hidden Markov Models (HMMs), Artificial Neural Networks (ANN), Dynamic Time Warping (DTW), Gaussian Mixture Models (GMM), etc. From these options, we have chosen an approach based on HMM as the model has a proven track record for many sound classification applications (Cowling, 2004; Okuno et al., 2004; Rabiner and Juang, 1993). Another advantage is that it can be easily implemented using the HMM Tool Kit (HTK) (Young, 1995), which is an open source Toolkit available on the Internet. It is also notifying that HTK was originally designed for speech recognition, which means we need to adapt the approach when applying it for environmental sounds of our interest.

Each sound file, corresponding to a sample of a sound type, was processed in frames pre-emphasized and windowed by a Hamming window (25 ms) with an overlap of 50%. A feature vector consisting of a 39-order MFCC characterized each frame. We modeled each sound using a left-to-right forty-state continuous-density HMM without state skipping.

Each HMM state was composed of two Gaussian mixture components. After a model initialization stage was done, all the HMM models were trained in eight iterative cycles.

4.2.3 Classifier

It was obvious that simple frequency characterization would not be robust enough to produce good classification results. To find representative features, previous study (Cowling, 2004) carried out an extensive comparative study on various transformation schemes, including the Fourier Transform (FT), Homomorphic Cepstral Coefficients (HCC), Short Time Fourier Transform (STFT), Fast Wavelet Transform (FWT), Continuous Wavelet Transform (CWT) and Mel-Frequency Cepstral Coefficient (MFCC). It was concluded that MFCC might be the best transformation for non-speech environmental sound recognition. A similar opinion was also articulated in (Okuno et al., 2004; Eronen et al., 2003). These finding provide the essential motivation for us to use MFCC in extracting features for environmental sound classification.

For classification, continuous HMM recognition is used. The grammar used is as follows:
(<alarm_clock | ambulance | body_spray | chop_board | cycle_bell | dish_cleaning | egg_fry | face_wash | female_cough | toilet_flush | food_eating | gargle | hair_dry | in_open_air_train | insect_semi | in_shower | in_subway_train | in_train_station | liquid_stir | lunch_time_cafe | male_cough | male_male_Speak | male_sniff | meeting_talk | microwave_oven | paper_flip | phone_vibration | plate_clutter | plate_knife | silence | pan_spoon | tap_water | tooth_brush | train_enter_leave | tv_watching | urination | vacuum_cleaner | vehicle_pass | water_basin | water_sink>), which means that there is no predefined sequence for all the activities and each activity may be repeated at time at any sequence.

4.2.4 Sound cues

The present system deals with 114 types of sounds that are grouped into 40 sound cues to infer 17 kinds of activities. For example, we have sound samples named as the following types, “eating soba”, “food munching”, “belching”, “spoon plate friction”, which are grouped into one sound cue called “food_eating”. Similarly, the sound samples named as, “news on TV”, “talk-show on TV”, “baseball game on TV”, “football game on TV” are grouped into “tv_watching” sound cue. By such grouping of the samples into a particular sound cue, we mean that the group of samples is used to train the classifier to recognize that particular sound cue. In this manner the present system is limited to 40 sound cues.

4.2.5 Object mapping

Each of the sound cue is annotated with a list of objects considering that those objects are conceptually connected or related to that particular sound cue. For example, Table 2 provides a set of sound cues along with the associated list of objects pertaining to the corresponding sound cue. After the classifier has detected an input sound sample to a particular sound cue, the “object mapping” module returns a list of associated object that are mapped to that particular sound cue. This list of object resolved from the sound cues are given to the common-sense knowledge module to facilitate activity inference.

4.2.6 Commonsense knowledgebase

Once the system gets the list of objects pertaining to the recognized sound cues, it calculates the probability of each object involved to the activities of our interest. For example, the

Sound Cue	Mapped Objects
vehicle_pass	road, bus, car, cycle, wind, people
toilet_flush	water, toilet, toilet-flush
in_shower	bathroom, shower, towel, water, body-wash
in_open_air_train	train, journey, people, announcement
food_eating	soba, noodle, people, chop stick, soup, rice, plate, cutleries, food, kitchen, restaurant
TV_watching	TV, game, living room, news
train_enter_leave	people, train, announcement, station
water_basin	restroom, water basin, wash, hand soap
water_sink	kitchen, sink, water, dish washer

Table 2. Sample list of sound classes and mapped objects

activity "eating" usually involves objects like food, plate, people and water. Requiring humans to specify the probabilities of involvement would be time consuming and difficult. Instead, an automatic approached is developed so that the system determines the probabilities by utilizing a technique adopted from Semantic Orientation (SO) (Hatzivassiloglou and McKeown, 1997; Grefenstette et al., 2004), employing NEAR search operator of AltaVista's web search result. The technique is described briefly as follows:
List of objects, O = {O_1, O_2, … O_K} (K=63)
List of locations, L = {L_1, L_2, …L_M} (M=9)
List of activities, A = {A_1, A_2, … A_N} (N=17)
Each location is represented by a set of English synonym words. WL_i = {W_1, W_2, … , W_P}. For example, L_1= "kitchen" and it is represented by, $W_{kitchen}$ = {"kitchen", "cookhouse", "canteen", "cuisine"}
$SA(O_i \mid L_j)$ = Semantic Associative value representing the object O_i to be associated with location L_j
$SA(O_i \mid A_j)$ = Semantic Associative value representing the object O_i to be associated with activity A_j
The formulae to get the *SA* values are indicated by Equation 1 and 2.

$$SA(O_i \mid L_j) = \log_2\left(\frac{\prod_{W \in WL_j} hits(O_i \text{ NEAR } W)}{\prod_{W \in WL_j} \log_2(hits(W))}\right) \quad (1)$$

$$SA(O_i \mid A_j) = \log_2\left(\frac{hits(O_i \text{ NEAR } A_j)}{\log_2(hits(A_j))}\right) \quad (2)$$

The obtained values support the concept that if an activity name and location co-occurs often with some object name in human discourse, then the activity will likely involve the object in the physical world. An excerpt from the list of SA values is provided in Figure 3.
Our approach is in the spirit of such manner while we use these obtained values as commonsense knowledgebase to assign a semantic associative value to the object pertaining to a sound sample as a model of relatedness in human activity. Thus, for example, if the system detects that the sound samples represent frying, saucepan, water sink, water, and

Object/Concept	Living Room	Work Place	Kitchen	Toilet	Gym	Inside Train	Train Station	Public Place	On Road	Max	Main Location
alarm	13.4523	10.9779	15.6387	14.3182	13.5427	13.6354	12.5480	14.7602	26.5531	26.5531	On Road
ambulance	10.6623	9.8535	10.3223	10.4926	9.0773	13.2013	11.7684	13.9239	28.2662	28.2662	On Road
announcement	14.3604	11.1248	16.0972	9.2678	10.5512	13.8032	26.1961	17.9592	12.9867	26.1961	Train Station
basin	12.7283	5.9383	16.1699	28.9758	8.7026	11.7622	10.0058	13.9179	16.5886	28.9758	Toilet
bicycle	12.0660	11.0806	13.8415	12.0582	13.3464	15.3610	13.5217	15.2338	32.6992	32.6992	On Road
blender machine	11.8717	3.0209	13.9110	9.2424	9.4124	8.9493	5.4949	7.5778	8.4731	13.9110	Kitchen
boiling	9.5302	6.1547	24.4261	9.4636	6.5960	9.6467	6.4360	11.0327	20.3290	24.4261	Kitchen
bottle	13.0488	9.5051	26.6152	14.2667	13.1362	14.0160	11.7199	13.2167	19.7792	26.6152	Kitchen
bowl	13.7426	9.8700	17.8777	14.2508	12.7417	13.9041	11.0942	14.3702	16.6401	17.8777	Kitchen
bus	14.2397	12.9289	14.6048	13.9183	14.3850	20.2331	19.0066	17.8408	35.9963	35.9963	On Road
car	17.1100	14.6828	18.1602	16.0323	16.8431	20.0676	18.1988	18.4553	39.7680	39.7680	On Road
cash register	12.1217	9.9955	11.9475	9.4562	9.6013	11.4610	9.5857	25.1415	17.6097	25.1415	Public Place
CD player	25.6021	9.1017	16.1107	12.8777	12.4728	12.5130	10.2689	13.8115	15.8466	25.6021	Living Room
chopping board	8.5498	3.8075	14.2755	7.0313	4.6498	9.3692	8.6712	11.5121	7.8466	14.2755	Kitchen
computer	16.3240	31.8790	17.5224	14.1969	15.2711	16.4552	13.8168	17.9183	23.2398	31.8790	Work Place
computer keyboard	12.3061	18.7801	10.6458	2.6458	10.2165	11.9073	9.1364	11.6065	14.6086	18.7801	Work Place
computer mouse	13.2372	23.1333	14.8366	9.0008	9.6654	11.8587	9.7082	10.4672	13.3829	23.1333	Work Place
cracking	10.6623	9.0055	11.8475	9.3253	8.5748	10.3801	8.2313	12.0505	13.3048	13.3048	On Road
drawer	12.9415	7.9088	22.7453	11.3453	9.7799	12.5932	9.0008	10.9119	12.3491	22.7453	Kitchen
dumbbell	3.7768	0.3912	5.6062	2.7826	13.1475	8.9878	4.0432	4.0513	7.5435	13.1475	Gym
electric train	11.7125	9.1475	11.3674	12.0942	10.6748	21.0071	25.2866	19.5850	20.0099	25.2866	Train Station
electric train door	12.4523	5.1316	13.4495	11.7333	9.6903	20.9529	16.2980	13.1332	12.5128	20.9529	Inside Train
escalator	6.2726	4.2041	8.3290	8.1634	7.7694	12.6316	12.1842	22.2830	15.5998	22.2830	Public Place
exhaust fan	10.6772	5.5149	19.1245	11.4133	6.1127	9.4541	6.4716	10.2535	9.2111	19.1245	Kitchen
female voice	12.8120	10.0351	12.0059	10.2803	9.3277	12.6765	11.5340	14.7703	13.0861	14.7703	Public Place
flush	11.4644	8.6960	14.3849	23.2391	9.1508	10.6882	8.9278	12.9179	17.8435	23.2391	Toilet
free weights	9.9000	5.6122	11.2942	8.2078	21.1491	13.4281	9.2604	11.1668	12.1073	21.1491	Gym
frying	8.6380	2.7667	16.6655	7.0626	4.5181	9.4947	4.3209	8.8550	11.5229	16.6655	Kitchen
gas	17.6590	13.1532	21.5666	15.1191	13.2346	15.7500	14.0495	17.6987	19.1024	21.5666	Kitchen
glass mug	16.3417	12.7021	30.9630	15.6146	13.5848	15.0419	13.2498	15.1203	26.0057	30.9630	Kitchen
hot drink	13.3394	9.5398	25.4527	13.7928	12.0367	12.9505	11.4037	14.1870	20.6605	25.4527	Kitchen

Fig. 3. Semantic Associative (SA) values for objects and locations of our interest

chopping board from consecutive input samples the commonsense knowledge usually infers a cooking activity located in kitchen.

Moreover we have also considered another kind of knowledge namely, ontology of daily-life that contains usual routines of peoples of different roles in terms of their presence at specific locations at specific time. In this initial prototyping three types of users are considered and based on empirical study the ontology of daily life (listed in Table 3) of each type of user is considered along with common sense knowledge base to perform legitimate inference. For example, if the system detects a sound cue related to eating activity during BN (before noon) time frame on a weekday and if the user is a service-holder, it doesn't log the event because the user is not happened to be present in a kitchen around that time. We also plan to provide the flexibility to personalize such ontology of daily life on the basis of each respective user.

Role	Weekday							Weekend						
	EM	M	BN	AN	E	N	LN	EM	M	BN	AN	E	N	LN
Graduate Student	H	H, T	H, U, K,T, R, Tr	U, G, T, P	H, U, R, Tr, K, T, G, P	H, U, R, Tr, K, T, P	H, K, T	H	H, T	H, K,T, R, Tr	G, T, P	H, R, Tr, K, T, G, P	H, R, Tr, K, T, P	H, K, T
Service-Holder	H	H, K, T, R, Tr	O, T	O, T	O, T, H, R, Tr, P	H, P, R, Tr, T	H	H	H, K, T	H, T, P, R, Tr	H, T, P, R, Tr, G	H, T, R, Tr, P	H, P, R, Tr, T	H
Elderly People	H, T	H, T, R, K	H, T, R, Tr, P	H, T, R, Tr, P	H, T, R, Tr, K, P	H	H,T	H, T	H, T, R, K	H, T, R, Tr, P	H, T, R, Tr, P	H, T, R, Tr, K, P	H	H,T

H: Home; O: Office; U: University; K: Kitchen; T: Toilet; G: Gym; R: Road; Tr: Train; P: Public place EM: Early Morning [03:00 – 05:00], M: Morning [05:00 – 8:00], BN: Before Noon [8:00 – 12:00], AN: After Noon [12:00 – 17:00], E: Evening [17:00 – 20:00], N: Night [20:00 – 01:00], LN: Late Night [01:00 – 03:00]

Table 3. An example of Ontology of daily life

4.2.7 Inference engine

The system continuously listens to the environment but it records sounds for ten seconds with an interval of ten seconds pause between two recordings. Thus for a minute the system gets three sound clips of equal length (i.e., ten seconds) that serves as the input to the classifier to get three sound cues in one minute. After that the object-mapping module provides a list of objects pertaining to the recognized sound cues. In this manner, the system produces a list of objects at every minute. The inference engine works by considering the list of objects that are prepared in every three minutes of time. This list of objects is then consulted with the involvement probabilities of activities stored in the commonsense knowledge base. An example of inference process is described in the sub-section "A Schematic Solution".

4.2.8 Script generator

At a preset specific time (e.g., midnight of each day) the whole day's activities are summarized and a daily report is automatically generated and archived at a database for further references. This report typically contains a consolidation of the frequency of occurrences of major activities of interest. A series of pre-prepared words are used and intelligently strung together to form simple sentences that conform to the basic rules of English grammar. An example of this is given as following:

Daily Report of Mr. Rahman's Activity, December 12, 2008

Mr. Rahman woke up at 8:20 am in the morning. He went to bathroom then. He had his breakfast at 9:38 am. He spent most of his time in living room. He watched TV. He had lunch at 2:13 PM. He talked with people. He went out and walked by the roads during evening. He didn't take shower today. At night he talked with someone.

The generated script is then automatically emailed to preset email address. The archival of these daily reports may enable a caregiver or a doctor to review records very quickly and in the process, build a detailed understanding of the subject's daily behavioral/activity patterns.

4.3 A schematic solution

The system continuously listens to the environment and captures sound regularly at 10 seconds of interval for 10 seconds of duration. Each captured sound-clip is sent to a computer wirelessly where the clip is processed. Thus for a three-minute interval the system gets nine sound clips. These nine sound clips are considered to infer an activity at that moment. Each sound clip is processed by a HMM classifier that outputs a particular sound cue. Each sound cue actually represents some real-word objects out of which interaction the sound maybe produced. For example, Let's assume that the system received nine sound clips to process and the HMM classified those nine clips into the following five unique sound cues: "chop_board", "pan_spoon", "water_sink", "plate_clutter", and "plate_knife". These five classes of sounds are pre-mapped to the objects as following,
"chop_board" → {knife, chop-board, onion, meat, people}
"pan_spoon" → {cooking-pan, spoon, stirrer, people}
"water_sink" → {water-sink, water, dish-washer}

"plate_clutter" → {cup-board, plate, glass, saucepan, people}
"plate_knife" → {plate, knife, spoon, people}
The list of interacting objects is consulted with the Semantic Associative (SA) value of the activities and locations stored in the commonsense knowledgebase. From this example, the unique list of objects obtained from the sound cues, U= {chop-board, cooking-pan, cup-board, dish-washer, glass, knife, meat, onion, people, plate, saucepan, spoon, water-sink, water, stirrer}. In this case, the objects yield a maximum SA value of having a relationship with "cooking" activity in "kitchen" location and the near candidates are "eating", "drinking tea/coffee" as activities. From the ontology of daily life the system finds that it is likely to have "cooking" activity at the concerned time (i.e., Table 3). Therefore, the activity inference engine considers this event as a legitimate event and logs it by either animating this event on a virtual world or generating a text message. This schematic solution is depicted in Figure 4.

5. Development challenges

In order to develop this system we have the following challenges and concerns:

5.1 Environmental sound corpus and features

For simplicity our collected sound corpus is limited to sound cues related to certain genres like, cooking, bathroom activity, human physiological action, outdoor, home, etc. Since the recognition rate is the only performance measure of such system, it is hard to judge how suitable the selected feature is. To solve this problem, it is essential to analyze the feature itself, and measure how good the feature itself is. It is concluded that MFCC may be the best transformation for non-speech environmental sound recognition (Okuno et al., 2004). This finding provides the essential motivation for us to use MFCC in extracting features for environmental sound classification.

5.2 Computing power of iPhone or Hand Held Devices

The computing potentials of HHDs are increasing in a rapid manner with respect to memory and incorporation of full-fledged operating system but they are still inferior to personal computer comparing to the speed of data processing. Running of the core signal-processing task requires high-end computing and therefore the processing should be done by means of external devices connected to the HHDs through the Bluetooth or wireless data transmission. Therefore in this case a light process to capture environmental sounds after some intervals will be running on the HHDs to be transmitted wirelessly to a server for activity detection.

5.3 Privacy and social issues

To manage their exposure, users should be able to disconnect from the virtual worlds at any time or be able to interrupt the sound capturing of their activities. The actual representation of users in the virtual world may be disclosed according to pre-configured user policies or users list. Additional privacy issues are related to the collection of environmental data by means of people carrying devices to different places and data collected about them. We consider important privacy and security challenges related to people-centric sensing (Mihailidis et al., 2001; Wan 1999; Barger et al., 2002).

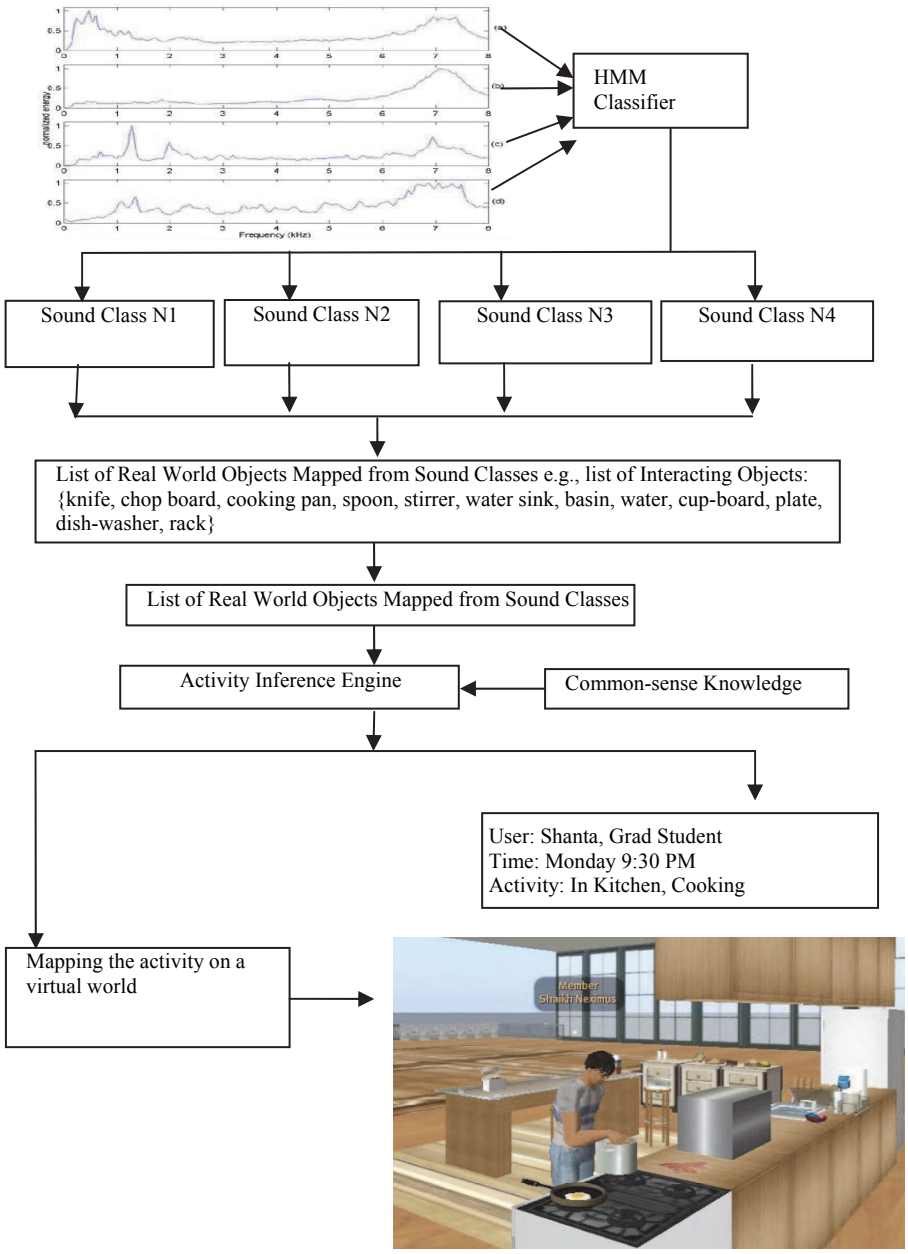


Fig. 4. An schematic solution of our approach

5.4 Scalability of the solution

Our default approach is to run activity recognition algorithms on the mobile HHDs to decrease communication costs and also to reduce the computational burden on the server. However, we recognize that signal processing based classification algorithms may be too computationally intensive for present HHDs and propose to run the classifier on a back-end server in this case. This may be particularly appropriate during the training phase of a classification model.

6. Experiment results and discussion

The purpose is to test the performance of the system in recognizing the major activities of our interest. The system was trained and tested to recognize the following 17 activities: Listening Music, Watching TV, Talking, Sitting Idle, Cleaning, Sitting idle, Working with PC, Drinking, Eating, Cooking, Washing, Urinating, Exercising, Waiting for Train, Traveling by Train, Shopping, Traveling on Road. As explained earlier, the sound samples recording for each activity was carried out separately. For example, for Listening Music, each subject played a piece of music of his/her choice, with this repeated a number of times for the same individual. The other subjects followed the same protocol and the entire process was repeated for developing the sound corpus for each activity being tested. It is also noted that we have various kinds of sounds (i.e., different sound-clip types) that are grouped together to represent one particular activity. The training data set was formed utilizing a 'leave-one-out' strategy. That is, all the samples would be used for their corresponding models' training except those included in the signal under testing. Hence, each time the models were trained respectively to ensure that the samples in the testing signal were not included in the training data set.

Since each sound clip resolves to a set of objects pertaining to the recognized sound class which is then considered to infer activity and location related to that sound class, we developed perceptual testing methodology to evaluate the system's performance on continuous sound streams of various sound events to infer location and activity. 420 test signals were created, each of which contained a mixture of three sound clips of respective 114 sound types. Since these 420 test signals are the representative sound clues for the 40 sound classes to infer 17 activities, we grouped these 420 test signals into 17 groups according to their expected affinity to a particular activity and location. Ten human (i.e., five male, five female) judges were engaged to listen to the test signals and judge an input signal to infer the activity from the given list of 17 activities (i.e., forced choice judgment) as well as the possible location of that activity from the list of given nine locations of our choice. Each judge was given all the 17 groups of signals to listen and assess. The number of test signals in a group varied from 3 to 6 and each test signal was the result of three concatenated sound clips of same sound type. Therefore a judge had to listen each test signal to infer the location and activity that the given signal seemed most likely to be associated with. In the same way the signals were given to the system to process. For the system the entire group of signals was given at a time to output one location and activity for each group. Since human judges judged each signal individually, in order to compare the result with the system, a generalization on the human assessment was done. The generalization was done in the following manner. A group of signals had at lest more than 3 signals and each of the signals was assigned a location and activity label by the judges. Thus a group of signals obtained a list of locations and activities. We counted the frequencies of location and activity labels for

each group assigned by each judge and took the maximum of the respective labels to finally assign the two types of labels (i.e., activity and location) for the group of signals. For each type of label, if more than one labels obtained equal frequency the random choice of the labels are considered. Thus we considered the judges' labels and system's inference with respect to the expected labels for the 17 groups of signals. Recognition results for activity and location are presented in Figure 5 and 6 respectively.

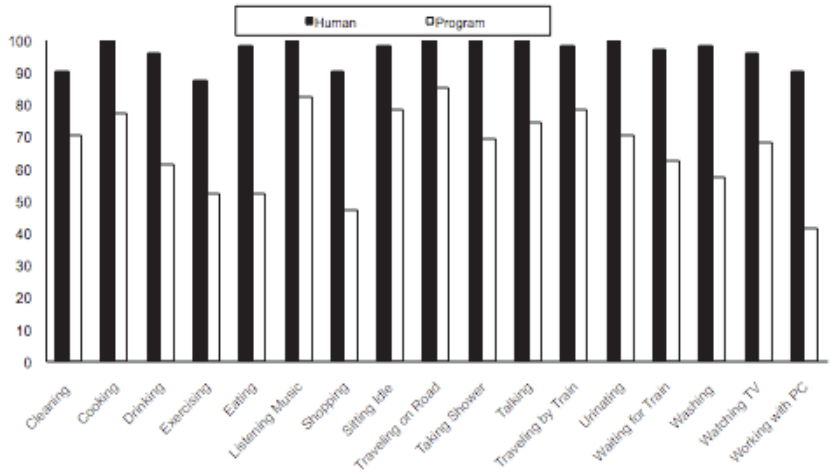


Fig. 5. Comparisons of recognition rates for 17 activities of our interest with respect to human judges

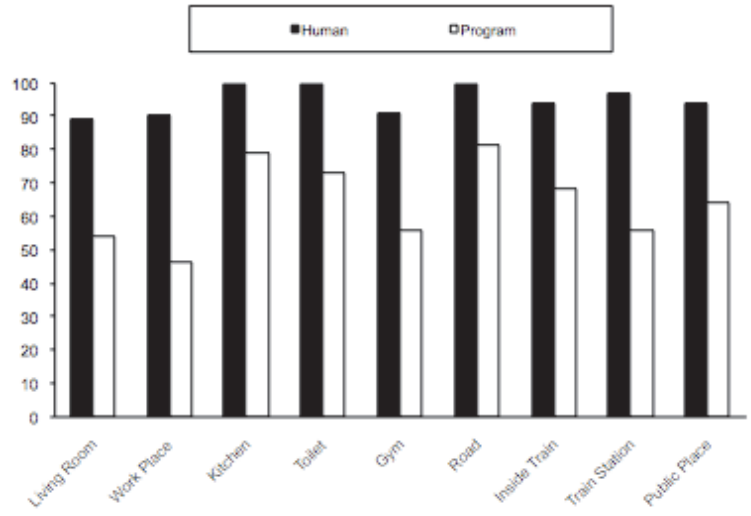


Fig. 6. Comparisons of recognition rates for 9 locations of our interest with respect to human judges

The recognition accuracy for activity and location is encouraging with most being above than 66% and 64% respectively. From Figure 5 and 6, we notice that humans are skillful in recognizing the activity and location from sounds (i.e., for humans' the average recognition accuracy of activity and location is 96% and 95% respectively). It is also evident that the system receives the highest accuracy (i.e., 85% and 81% respectively) to detect "traveling on road" activity and "road" location respectively, which is a great achievement and pioneer effort in this research that no previous research attempted to infer outdoor activities with sound cues. The correct classification of sounds related to activity "working with pc" and location "work place" were found to be very challenging due to the sounds' shortness in duration and weakness in strength, hence the increased frequency for them to be wrongly classified as 'silence' type sound class.

7. Conclusion

In this chapter, we described a novel acoustic indoor and outdoor activities monitoring system that automatically detects and classifies 17 major activities usually occur at daily life. Carefully designed HMM parameters using MFCC features are used for accurate and robust sound based activity and location classification with the help of commonsense knowledgebase. Experiments to validate the utility of the system were performed firstly in a constrained setting as a proof-of-concept and in future we plan to perform actual trials involving peoples in the normal course of their daily lives to carry the device running our application that listens to the environment and automatically logs the daily event based on the mentioned approach. Preliminary results are encouraging with the accuracy rate for outdoor and indoor sound categories for activities being above 67% and 61% respectively. We sincerely believe that the system contributes towards increased understanding of personal behavioral problems that significantly is a concern to caregivers or loving ones of elderly people. Besides further improving the recognition accuracy, we plan to enhance the capability of the system to identify different types of human vocalization, which provides useful information pertaining to the mental wellbeing of the subject. We also believe that integrating sensors into the system will also enable acquire better understanding of human activities. The enhanced system will be shortly tested in a full-blown trial on the most needy elderly peoples residing alone within the cities of Tokyo evaluating its suitability as a benevolent behavior understanding system carried by them.

This type of application may have another potential use. Software like Google Earth and Microsoft Virtual Earth map the real world with great accuracy in terms of geographical locations and local features. We believe that the next step is to enable a user to represent real world activities in this kind of virtual world whereby personal avatars will be able to reflect what the real persons are doing in the real world by inferring activities from sound cues. This type of application is a source of fun for young generation while it has lots of potential regarding virtual shopping mall in e-commerce context, easy monitoring for elderly people for the caregivers etc.

8. References

Chen, J; Kam, A. H., Zhang, J., Liu, N., and Shue, L. (2005). Bathroom Activity Monitoring Based on Sound, Proc. 3rd Int'l Conf, PERVASIVE 2005, Munich, Germany, LNCS 3468/2005, pp. 47-61.

Philipose, M ; Fishkin, K. P., Perkowitz, M., Patterson, D. J., Fox, D., Kautz, H., & HahnelInferring, D. (2004). Activities from Interactions with Objects. *IEEE Pervasive Computing*, Vol. 3, No. 4, pp. 50-57.

Temko, A. & Nadeu, C. (2005). Classification of meeting-room acoustic events with Support Vector Machines and Confusion-based Clustering. Proc. ICASSP'05, pp. 505-508.

Wang, D. & Brown, G. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press

Mihailidis, A. ; Fernie, G. & Barbenel, J.C. (2001). The Use of Artificial Intelligence in the Design of an Intelligent Cognitive Orthosis for People with Dementia. *Assistive Technology*, Vol. 13, No. 1, pp. 23–39.

Wan, D. (1999). Magic Medicine Cabinet: A Situated Portal for Consumer Healthcare. Proc. 1st Int'l Symp. Handheld and Ubiquitous Computing (HUC 99), LNCS 1707, Springer-Verlag, pp. 352–355.

Barger, T. ; Alwan, M., Kell, S., Turner, B., Wood, S., Naidu, A. (2002). Objective remote assessment of activities of daily living: Analysis of meal preparation patterns. Poster presentation, Medical Automation Research Center, University of Virginia Health System.

Tran, Q. ; Truong, K., & Mynatt, E. (2001). Cook's Collage: Recovering from Interruptions. demo at 3rd Int'l Conf. Ubiquitous Computing (Ubi-Comp 2001).

Glascock, A. & Kutzik, D. (2000). Behavioral Telemedicine: A New Approach to the Continuous Nonintrusive Monitoring of Activities of Daily Living. *Journal of Telemedicine*. Vol. 6, No. 1, pp. 33–44.

Korhonen, I. ; Paavilainen, P. & Särelä, A. (2003). Application of Ubiquitous Computing Technologies for Support of Independent Living of the Elderly in Real Life Settings. Proc. UbiHealth 2003: 2nd Int'l Workshop Ubiquitous Computing for Pervasive Healthcare Applications.

Mozer, M. (1998). The Neural Network House: An Environment That Adapts to Its Inhabitants. Proc. AAAI Spring Symposium. Intelligent Environments, tech. report SS-98-02, AAAI Press, pp. 110–114.

Campo, E. & Chan, M. (2002). Detecting Abnormal Behavior by Real-Time Monitoring of Patients. Proc. AAAI Workshop Automation as Caregiver, AAAI Press, pp. 8–12.

Guralnik, V. & Haigh, K. (2002). Learning Models of Human Behaviour with Sequential Patterns. Proc. AAAI Workshop Automation as Caregiver, AAAI Press, pp. 24–30.

house_n Project, http://architecture.mit.edu/house_n

Bush, V. (1945). As we may think, *Atlantic Monthly*, 1945

Rhodes, B. & Starner, T. (1996). Remembrance Agent: A Continuously Running Automated Information Retrieval System. Proc. 1st Int'l Conf. Practical App. of Intelligent Agents and Multi-Agent Technology, pp. 487-495.

Clarkson, B. & Pentland, A. (1999). Unsupervised Clustering of Ambulatory Audio and Video. Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, IEEE CS Press, Vol. 6, pp. 3037-3040.

Clarkson, B. ; Mase, K. & Pentland, A. (2001). The Familiar: A Living Diary and Companion. Proc. ACM Conf. Computer–Human Interaction, ACM Press, pp. 271-272.

Gemmell, J ; Bell, G., Lueder, R., Drucker, S. & Wong, C. (2002). MyLifeBits: Fulfilling the Memex Vision. Proc. ACM Multimedia, ACM Press, pp. 235-238.

Fitzgibbon, A. & Reiter, E. (2003). 'Memories for Life': Managing Information over a Human Lifetime. UK Computing Research Committee Grand Challenge proposal.

Vemuri, S. & Bender, W. (2004). Next-Generation Personal Memory Aids. *BT Technology Journal*, Vol. 22, No. 4, pp. 125-138.

Blum, M. ; Pentland, A. & Troster, G. (2006). InSense: Internet-Based Life Logging. , Vol. 13, No. 4, pp.40-48.

Dickie, C. ; Vertegaal, R., Fono, D., Sohn, C., Chen, D., Cheng, D., Shell, J.S. & Aoudeh, O. (2004). Augmenting and sharing memory with eyeBlog. Proc. 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, ACM Press.

Gemmell, J. ; Williams, L., Wood, K., Lueder, R. & Bell, G. (2004). Passive Capture and Ensuing Issues for a Personal Lifetime Store. Proc. Continuous Archival and Retrieval of Personal Experiences, ACM Press.

Raento, M. ; Oulasvirta, A., Petit, R. & Toivonen, H. (2005). ContextPhone: A Prototyping Platform for Context-aware Mobile Applications. *IEEE Pervasive Computing*, vol 4, no 2, pp. 51–59.

Panu, K.; Jani, M., Juha, K., Heikki, K. & Esko-Juhani, M. (2003). ContextPhone: Managing Context Information in Mobile Devices. *IEEE Pervasive Computing*, vol 2, no 3, pp. 42–51.

Kim, K.; Jung, M. & Cho, S. (2007). KeyGraph-based chance discovery for mobile contents management system Source. Int'l Journal of Knowledge-based and Intelligent Engineering Systems archive, vol 11, no 5, pp.313-320.

Cowling, M. (2004). Non-Speech Environmental Sound Recognition System for Autonomous Surveillance, Ph.D. Thesis, Griffith University, Gold Coast Campus.

Okuno, H.G.; Ogata, T., Komatani, K. & Nakadai, K. (2004). Computational Auditory Scene Analysis and Its Application to Robot Audition. International Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS), pp. 73–80

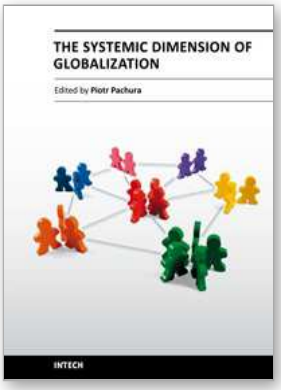
Eronen, A.; Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G. & Huopaniemi, J. (2003). Audio-based Context Awareness-Acoustic Modeling and Perceptual Evaluation. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), vol. 5, pp. 529-532.

Rabiner, L. R. & Juang, B.H. (1993). Fundamentals of Speech Recognition, PTR Prentice-Hall Inc, New Jersey.

Young, S. (1995). The HTK Book, User Manual, Cambridge University Engineering Department.

Hatzivassiloglou, V. & McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In 35th annual meeting on ACL, pp.174-181.

Grefenstette, G. ; Qu, Y. Evans, D. & Shanahan, J. (2004). Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In Computing Attitude and Affect in Text: Theory and Applications, eds. J. Shanahan, Y. Qu, and J. Wiebe, 93-107. The Information Retrieval Series Vol. 20, Netherlands: Springer Verlag.



The Systemic Dimension of Globalization

Edited by Prof. Piotr Pachura

ISBN 978-953-307-384-2
Hard cover, 288 pages
Publisher InTech
Published online 01, August, 2011
Published in print edition August, 2011

Today science is moving in the direction of synthesis of the achievements of various academic disciplines. The idea to prepare and present to the international academic milieu, a multidimensional approach to globalization phenomenon is an ambitious undertaking. The book The Systemic Dimension of Globalization consists of 14 chapters divided into three sections: Globalization and Complex Systems; Globalization and Social Systems; Globalization and Natural Systems. The Authors of respective chapters represent a great diversity of disciplines and methodological approaches as well as a variety of academic culture. This is the value of this book and this merit will be appreciated by a global community of scholars.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mostafa Al Masum Shaikh, Keikichi Hirose and Mitsuru Ishizuka (2011). Recognition of Real-World Activities from Environmental Sound Cues to Create Life-Log, The Systemic Dimension of Globalization, Prof. Piotr Pachura (Ed.), ISBN: 978-953-307-384-2, InTech, Available from: http://www.intechopen.com/books/the-systemic-dimension-of-globalization/recognition-of-real-world-activities-from-environmental-sound-cues-to-create-life-log

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821