NAIST-IS-DD1261015

Doctoral Dissertation

Human Activity and Environment Recognition on Mobile Devices

Yuki Maruno

January 30, 2015

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Yuki Maruno

Thesis Committee:

Professor Kazushi Ikeda	(Supervisor)
Professor Shoji Kasahara	(Co-supervisor)
Professor Keiichi Yasumoto	
Associate Professor Paul Pang	(Unitec Institute of Technology)

Human Activity and Environment Recognition on Mobile Devices[*]

Yuki Maruno

Abstract

mHealth, the use of mobile devices and other wireless technology in health care and public health, is a rapidly expanding area of research and practice. mHealth applications help people manage their own health, promote healthy living, and gain access to useful information. In building such applications, advanced mobile sensors are used. However, they consume too much energy.

In this dissertation, we propose methods toward mHealth applications, which deals with the limitation in available electric power. To reduce energy consumption, we use a single sensor for each of the following tasks, which are recognition of human activity and recognition of the user's environment. For human activity recognition, we use a three-axis accelerometer, which is equipped with almost any mobile device. In order to maintain high accuracy in recognition with low computational cost, we employ the wavelet transform and the singular value decomposition during feature extraction. For environment recognition on the other hand, we use a microphone to capture the environmental sounds. To ensure sufficient location coverage, the data collection is designed based on people's daily routines, which enables coverage of a wide range of environments including public transportation, offices, streets, and shopping malls. In order to classify the 17 environments, we make use of several audio features from time domain and frequency domain.

With regard to experimental results, the algorithm used was able to classify user activities into walking, running, standing still and being in a moving train

[*] Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1261015, January 30, 2015.

with accuracy of over 90%. As for environment recognition, accuracy of over 80% for the 17 environments was achieved. The proposed method deals with the limitation in available electric power, thereby addressing an mHealth application issue.

Keywords:

Context Awareness, Mobile devices, Accelerometer, Microphone, Wavelet transform, Singular value decomposition

モバイルデバイスを用いたユーザの行動および環境推定*

丸野 由希

内容梗概

近年，スマートフォンやタブレットなどのモバイル端末が急速に普及し，そのアプリケーションは多様化してきている．中でも，ユーザの状況（コンテキスト）に応じたサービス提供のためのアプリケーションが近年注目をあびている．それらのアプリケーションには，ユーザがおかれた状況や環境を知ることが重要である．

そこで本論文では，モバイル端末上で動作するアプリケーションのためのユーザ行動認識手法およびユーザの環境推定手法を提案する．技術の発展により，モバイル端末には，カメラ，GPS，加速度センサ，マイク，照度センサ，地磁気センサ，温度センサなど，様々なセンサが搭載されている．センサの多様化により，複数センサを組み合わせるなどして，より高度なユーザの状況推定が可能になった一方で，消費電力の問題が大きな課題となっている．例として，GPS による測位においては，通信により多大な電力を消費する．

本研究においては，ユーザ行動認識ではモバイル端末で一般的になっている単一の３軸加速度センサのみを使用し，低消費電力で高い認識精度の実現を試みた．特徴抽出には衝撃的な信号をよく抽出できるとされるウェーブレット変換を用いた．さらに特異値分解を用いて特徴選択を行い，ユーザ行動を「走行」,「歩行」,「静止」,「乗車中」の４状態に識別した結果，90%以上の識別率を実現した．一方，ユーザの環境推定では，モバイル端末に搭載されたマイクロフォンを使用して，公園やレストランといった，ユーザが現在おかれている環境を推定することを試みた．特徴量を工夫する事で，17 種類の環境において 80%以上の識別率を実現した．

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 博士論文, NAIST-IS-DD1261015, 2015 年 1 月 30 日.

キーワード

コンテキストアウェアネス，モバイル端末，加速度センサ，マイクロフォン，ウェーブレット変換，特異値分解

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The use of portable electronic devices such as mobile phones, tablet computers and personal digital assistants (PDAs) has been widely recognized as an efficient way to improve our daily life. In fact, most people always have their mobile devices with them [1]. These mobile devices are equipped with powerful embedded sensors, such as accelerometers, Global Positioning System (GPS), microphones, and cameras, which enable new applications across a wide range of domains, including business, healthcare, social networks, safety, environmental monitoring, and transportation [2]. The data from such sensors could be harnessed to provide valuable services or applications based on a user's situation or context without prompting the user, which is known as context awareness [3, 4]. For example, context aware devices could automatically turn on silent mode (no ring tone, but vibration) when the user is on a train. Context aware devices may also be used for health care purposes. If mobile devices record a user's context constantly, the information would be helpful when doctors give a diagnosis. It would also be practically useful for people to pay attention in time to their recent unhealthy behavior patterns and change them, such as adding more exercise, which could improve their health.

There have been various researches to estimate the user contexts from several sensor data. Iso et al. [5], for example, proposed a gait analyzer with an acceleration sensor on a mobile phone. They extracted features by the wavelet packet decomposition and classified them with a self-organizing algorithm based on Bayesian theory. Their algorithm could identify gaits such as walking, running, going up/down stairs and walking fast with an accuracy of about 80% in their experiments. Cho et al. [6] estimated the user contexts with a combination of acceleration sensor and GPS. They overcame the problem of confusion between standing still and being in a moving train in classifying walking, running, and the previous two contexts by using the GPS information in addition to the acceleration sensor. Although they achieved an accuracy of 90.6%, GPS sensors do not work well indoors or underground [7]. Moreover, high-load computation and using multiple sensors are energy consuming though a long standby time is another important aspect of a mobile phone.

1.1 Context Awareness in Healthcare

Context awareness in healthcare is a rapidly expanding area of research and practice. The use of mobile devices and other wireless technology in health care is commonly referred to as mHealth, which has the potential to change when, where, and how healthcare is provided. mHealth applications help people manage their own health and wellness, promote healthy living, and gain access to useful information. Apple and Google have recently announced their mHealth platforms successively. Apple introduced the HealthKit [8], an mHealth platform that allows health and fitness applications to share their data. Google developed Google Fit [9] for mHealth which allows various applications to share health data for individual users to create a complete picture of their fitness. Such mHealth platforms require sensing and inference to determine the user context.

In this dissertation, we propose methods toward mHealth applications, which deals with the limitation in available electric power. In building such applications, the following technical problems need to be solved: a) recognition of human activity and b) recognition of the user's environment. Instead of relying on advanced mobile sensors, we use a single embedded sensor, a three-axis accelerometer for human activity recognition and a microphone for human environment recognition, which are equipped with almost any mobile phone.

1.2 Organization of Dissertation

This dissertation is organized as follows. In Chapter 2, we propose a new approach to recognize human activity for mobile applications. We first introduce the components of the proposed method, the wavelet transform and the singular value decomposition, and then describe the details of the experiments as well as the results. We also discuss the mother wavelet and an alternative of the wavelet transform. In Chapter 3, we propose a new approach to recognize the user's environment for mobile applications. We first give a review of previous studies on environmental sound recognition, and then describe data collection and data labeling in multi-task. We discuss audio feature extraction for environmental sound classification and then analyze the environmental sound towards sound health understanding. Chapter 4 concludes this study with detailed discussions and recommendations for future research.

Chapter 2

Human Activity Recognition on Mobile Devices

2.1 Introduction

Mobile devices including mobile phones are daily necessities in modern society. Since modern mobile devices are equipped with multiple sensors such as microphones, cameras, Global Positioning System (GPS) and accelerometers, the data from such sensors could be harnessed to provide valuable services or applications based on user's context. In this chapter, we propose a method for mobile devices to recognize a user's context. In order to reduce the electric-power consumption, the method uses only a single three-axis accelerometer equipped with almost every mobile device. The signal of the accelerometer is transformed by the wavelet transform [10] since its effectiveness was shown by Mantyjarvi et al. [11] after a normalization so that the signal is device-direction-free. Here, we employ the Haar mother wavelet for a low computational load. The high dimensionality of the wavelet coefficients is reduced using the singular value decomposition (SVD) for a high accuracy and a low computational cost [12, 13]. The largest and the second largest singular values are used as the input of the classifier to four states: walking, running, standing still and being in a moving train. Our classifier is a multi-layer perceptron (MLP) [14].

2.2 Proposed Method

Our method classifies the user's contexts into four states, walking, running, standing still and being in a moving train, with a single three-axis accelerometer on a mobile device. Its flow is depicted in Fig. 2.1. The three-axis accelerometer outputs time-series of the X-, Y- and Z-axis accelerations. These are preprocessed to a direction-free sequence. Then, the wavelet transform extracts its features, which the SVD reduces to a two-dimensional signal at a time. Finally, the signal is classified by an MLP trained with learning data. The details are described in the following subsections.

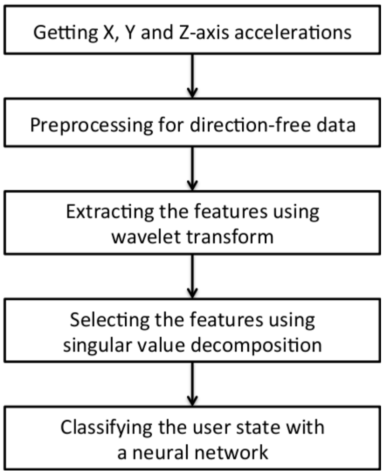

Getting X, Y and Z-axis accelerations
Preprocessing for direction-free data
Extracting the features using wavelet transform
Selecting the features using singular value decomposition
Classifying the user state with a neural network

Figure 2.1. Workflow of the proposed method

2.2.1 Preprocessing for Direction-Free Signals

In this study, we assume that a user carries a mobile device steadily but unrestrictedly in the direction. This means we consider the case where a mobile phone is in a pocket or a handbag. Fig. 2.2 shows examples of "standing still" data and "train" data. Although both Fig. 2.2 (a) and (b) are "standing still" data, they were measured at different device direction. Since the signal produced by the three-axis accelerometer in the device is three-dimensional time-series dependent on the direction of the device (Fig. 2.2 (a)(b)), it should be preprocessed so as to be device-direction-free. We chose the magnitude of the acceleration (Fig. 2.3) as a device-direction-free signal calculated from the three-dimensional signal, that is,

$$f(t) = \sqrt{X^2(t) + Y^2(t) + Z^2(t)}, \tag{2.1}$$

where $X(t)$, $Y(t)$ and $Z(t)$ are the values of X-, Y- and Z-axis accelerations at time t, respectively. Note that (a), (b), and (c) of Fig. 2.2 correspond to those of Fig. 2.3.

2.2.2 Feature Extraction

Cho et al. [6] reported that "standing still" and "train" were frequently confused due to their similar waveforms (Fig. 2.3) and statistics (Table 2.1). Note that (a), (b), and (c) of Table 2.1 correspond to those of Fig. 2.3.

Table 2.1. Basic statistics of Fig. 2.3.

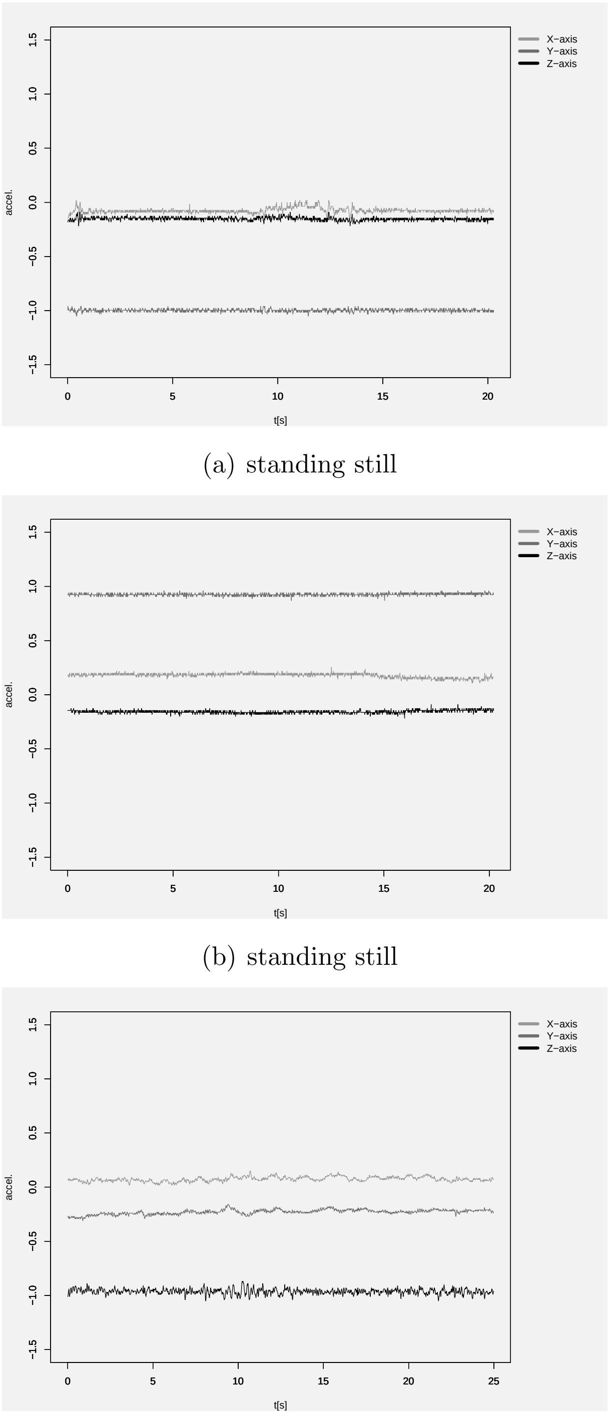
	Max	Average	Variance
(a) standing still	1.064	1.012	0.00013
(b) standing still	0.996	0.956	0.00012
(c) train	1.079	0.993	0.00054

To discriminate these states and two others, we employed the wavelet transform that is a tool for nonstationary time-series analysis [10]. The wavelet coefficient $W(a, b)$ of a time-series $f(t)$ with scale parameter a and translation parameter b is defined as

$$\begin{aligned} W(a,b) &= \langle f(t), \Psi_{a,b}(t) \rangle && (2.2) \\ &= \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{a}} \Psi^* \left(\frac{t-b}{a} \right) dt, && (2.3) \end{aligned}$$

where $\Psi(t)$ is a fixed function called the mother wavelet and * denotes the complex conjugate. We used the Haar mother wavelet

$$\Psi(t) = \begin{cases} 1, & \text{if } 0 \leq t < \frac{1}{2} \\ -1, & \text{if } \frac{1}{2} \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

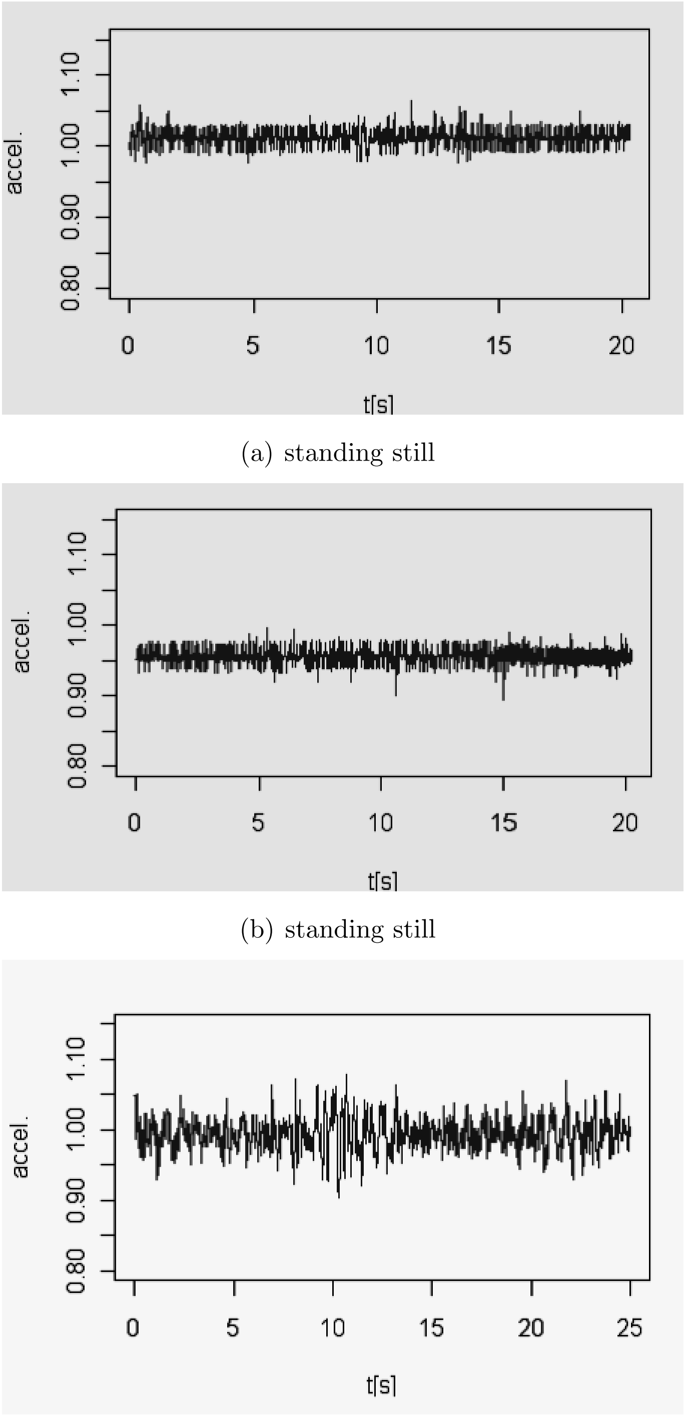


(a) standing still

(b) standing still

(c) train

Figure 2.2. Examples of data. X- (light gray), Y- (dark gray), Z- (black) axis acceleration of two “ standing still ” data (a, b) and one “ train ” data (c).



(a) standing still

(b) standing still

(c) train

Figure 2.3. The magnitude of the acceleration in Fig. 2.2.

since the wavelet coefficients are calculated without multiplications, which leads to a low computational load. See Fig. 2.4 for examples of the wavelet transform.

2.2.3 Dimensionality Reduction

In general, the wavelet coefficients are high-dimensional. For example, if window length is 100 and scale parameter is 55, wavelet coefficients are 5,500. High dimensionality leads to not only an electric-power consumption but also degradation of performance due to overfitting [15]. To overcome these problems, we reduced the dimensions using the singular value decomposition (SVD).

The SVD decompose an $m \times n$ real-valued matrix $X \in R^{m \times n}$ to $X = U \Sigma V^T$, where $U \in R^{m \times m}$ and $V \in R^{n \times n}$ are orthogonal matrices and $\Sigma \in R^{m \times n}$ is a diagonal matrix. The diagonal elements of Σ, $\sigma_1, \ldots, \sigma_{\min(m,n)}$ are called the singular values. By convention $\sigma_1 \geq \sigma_2 \geq \cdots$ is assumed.

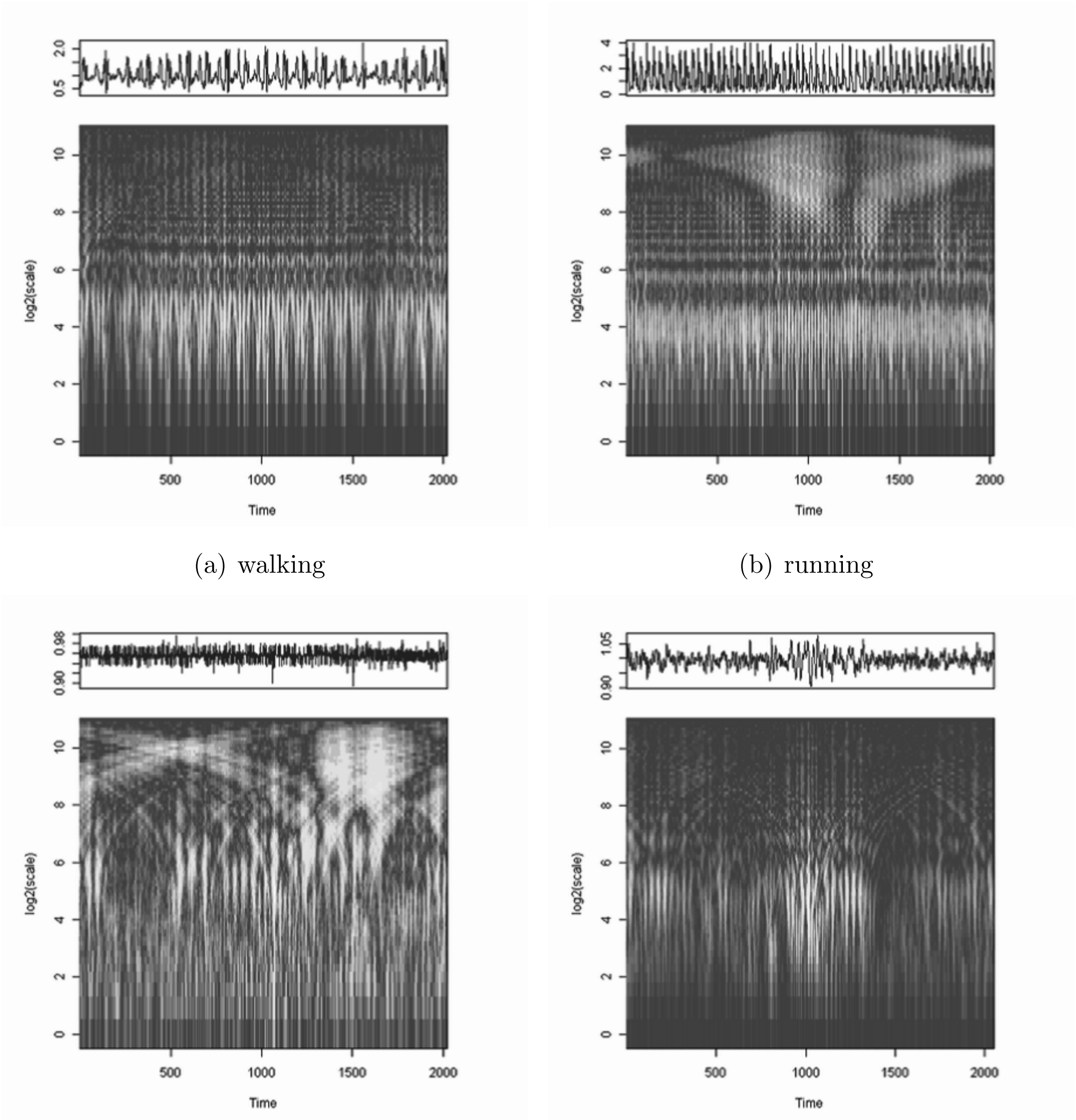
The conventional dimensionality reduction based on the SVD projects data into the subspace spanned by the k principal components for a fixed k. However, our method employed the singular values themselves. That is, the feature vector of our method was $(\sigma_1, \sigma_2, \ldots, \sigma_k)$. In fact, our preliminary experiments showed that the four states were well separated in the space of (σ_1, σ_2), that is, in the case of k=2 (Fig. 2.5).

2.2.4 Classification

The final procedure is to classify feature vectors to one of the four states, walking, running, standing still, and being in a moving train. Since the states were well separated in the space of (σ_1, σ_2), we employed the standard multi-layer perceptron (MLP) with five hidden nodes. We selected the number of hidden nodes based on our preliminary experiments. The MLP was trained using the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton method [16].

2.3 Experiments

In order to confirm the effectiveness of our method, the authors carried out some experiments for user-state recognition.



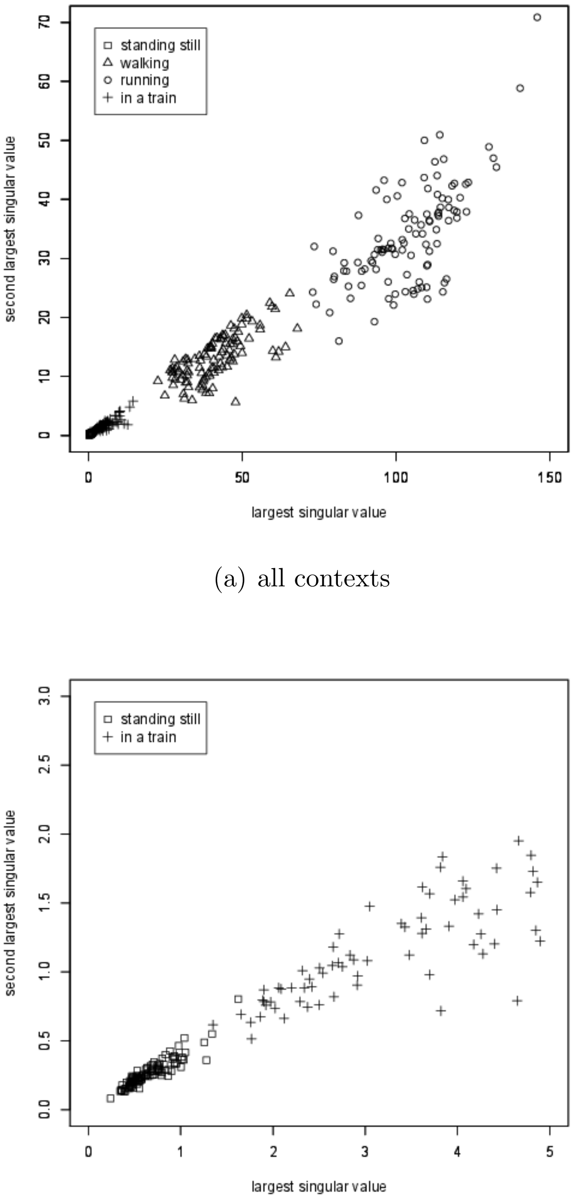
(a) walking

(b) running

(c) standing still

(d) being in a moving train

Figure 2.4. Examples of wavelet transform of the magnitude of the acceleration



(a) all contexts

(b) standing still and in a train (the enlarged figure of Fig. 2.5(a))

Figure 2.5. Plot of feature vectors in our preliminary experiments.

2.3.1 Materials

The data for walking, running and standing still were collected from HASC2010 corpus * while those for being in a moving train were originally measured since the corpus did not include such data. HASC2010 corpus originally include six activities: “stay”, “walk”, “jog”, “skip”, “stair up”, and “stair down”. In our research, we only used “stay”, “walk”, and “jog”, which corresponds to “standing still”, “walking”, and “running”, respectively. In general, signal data may vary over time or situation in real world practice. For example of moving train, signals may differ when person is standing or sitting in a train. Thus, several measurements were conducted at different situations and trains to ensure the situation coverage of data collection. Eight participants (seven in the corpus; one in our experiment) carried their mobile phones as they liked (Table 2.2). Each participants conducted the measurement five times for each activity, which means each activity has 35 data files. In total, 140 data files were used. We divided our dataset at random into training (20%) and test (80%) sets.

Table 2.2. Position of mobile phones

Participant No.	Sensor position
1	waist pocket or hand
2	waist pocket
3	waist pocket
4	breast pocket
5	waist pocket or hand
6	waist pocket or hand
7	upper arm (fixed)
8	bag or hand

2.3.2 Recognition Accuracy

Since the recognition accuracy and electric-power consumption increase as the sampling rate and the time-window width increase in general, we saw the de-

*http://hasc.jp/hc2010/HASC2010corpus/hasc2010corpus-en.html pendency of the recognition accuracy on the sampling rate and the time-window width.

The recognition accuracy of each condition was described in Table 2.3. We used F-measure for our evaluation. First, we calculated the Precision and Recall defined as:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{2.5}$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{2.6}$$

where TP, FP and FN mean True Positive, False Positive and False Negative, respectively. Then, F-measure for each context was calculated defined as

$$F = (2\text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision}). \tag{2.7}$$

F-measure in Table 2.3 is mean value of four contexts.

Our method achieved an accuracy of more than 90 % if the sampling rate is more than 25 Hz and the time-window width is more than 1 second. The accuracy of our method is better than the conventional method in [6] except three conditions where time window of the wavelet transform is less than 1 second (Fig. 2.6).

Table 2.3. Average Accuracy

Rate \Width	0.5 sec	1 sec	2 sec	3 sec
10 Hz	84.9%	88.1%	90.7%	91.8%
25 Hz	89.2%	92.6%	92.5%	92.5%
50 Hz	90.5%	92.9%	94.1%	93.0%
100 Hz	91.0%	93.9%	93.6%	93.6%

2.3.3 Computation Time

Since the electric-power consumption is proportional to the computational complexity, we evaluated the computation time instead. The authors performed our method and the conventional method in [6] with R on a PC with Intel XEON(R) 3.20 GHz.

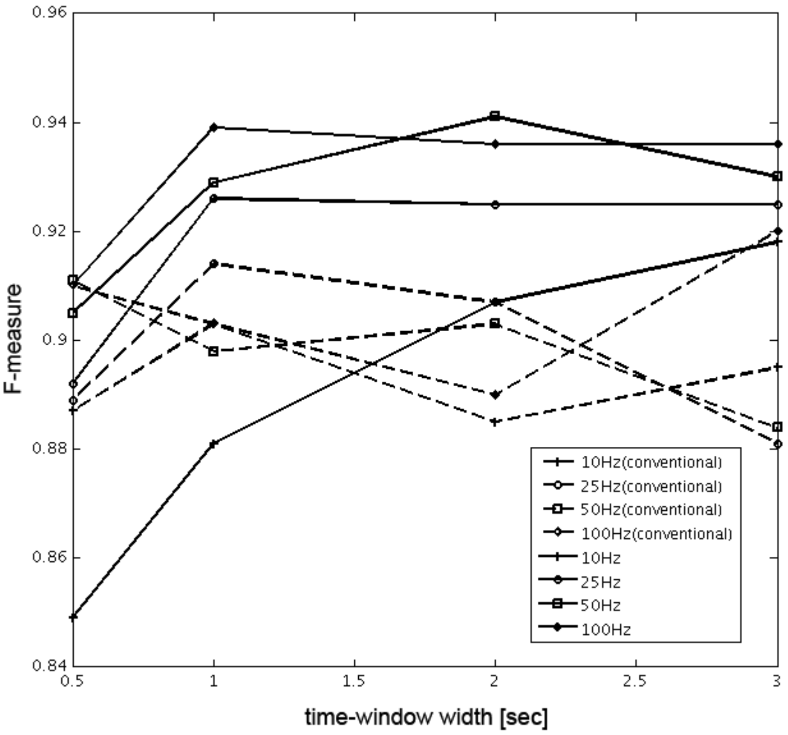


Figure 2.6. Accuracy comparison of our method (solid lines) with the conventional method (dashed lines).

The results of our method and the conventional method show that the computation time of our method increases in the order of O(N) and that it is only twice as much as that of the conventional method when the sampling rate is 10 Hz (Fig. 2.7). This implies that our method is available to smart devices since the conventional method has already worked on smart devices and the hardware performance of smart devices improves according to Moore's Law [17]. In fact, the Geekbench 3 scores of iPhone 4, iPhone 5 and iPhone 6 are 207, 710 and 1610, respectively [18].

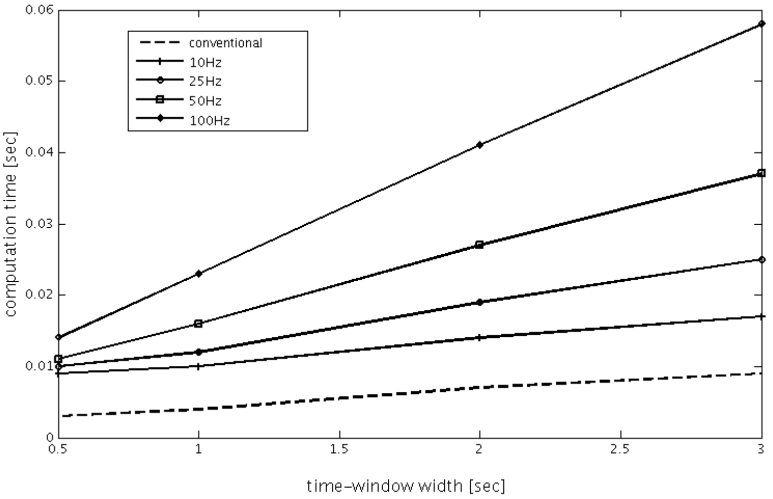


Figure 2.7. Computation time comparison of our method (solid lines) with the conventional method (dashed line).

2.4 Discussion

In our experiments, our method performed better than and had twice as much computation time as the conventional method. Since the conventional method uses the GPS [6], it might consume more electric power than ours.

2.4.1 Mother Wavelet Comparison

We used the Haar mother wavelet because of its computation load. However, there are several mother wavelets such as Gaussian and Mexican hat. We compared them in terms of accuracy and computation time.

Tables 2.4 and 2.5 show the accuracy for each mother wavelet and the calculation time per estimation, respectively. Although the accuracy is almost the same, the calculation time of Haar mother wavelet is less than half of the others, indicating that the Haar mother wavelet is suitable for our method.

Table 2.4. Dependency of accuracy on mother wavelet

Time-window width	0.5s	1s	2s	3s
Haar	91.0%	93.9%	93.6%	93.6%
Mexican hat	91.1%	94.3%	93.9%	93.9%
Gaussian	91.2%	94.1%	93.5%	94.1%

Table 2.5. Dependency of computation time on mother wavelet

Time-window width	0.5s	1s	2s	3s
Haar	0.014sec	0.023sec	0.041sec	0.058sec
Mexican hat	0.029sec	0.062sec	0.129sec	0.202sec
Gaussian	0.029sec	0.061sec	0.128sec	0.200sec

2.4.2 Comparison with Short-Time Fourier Transform

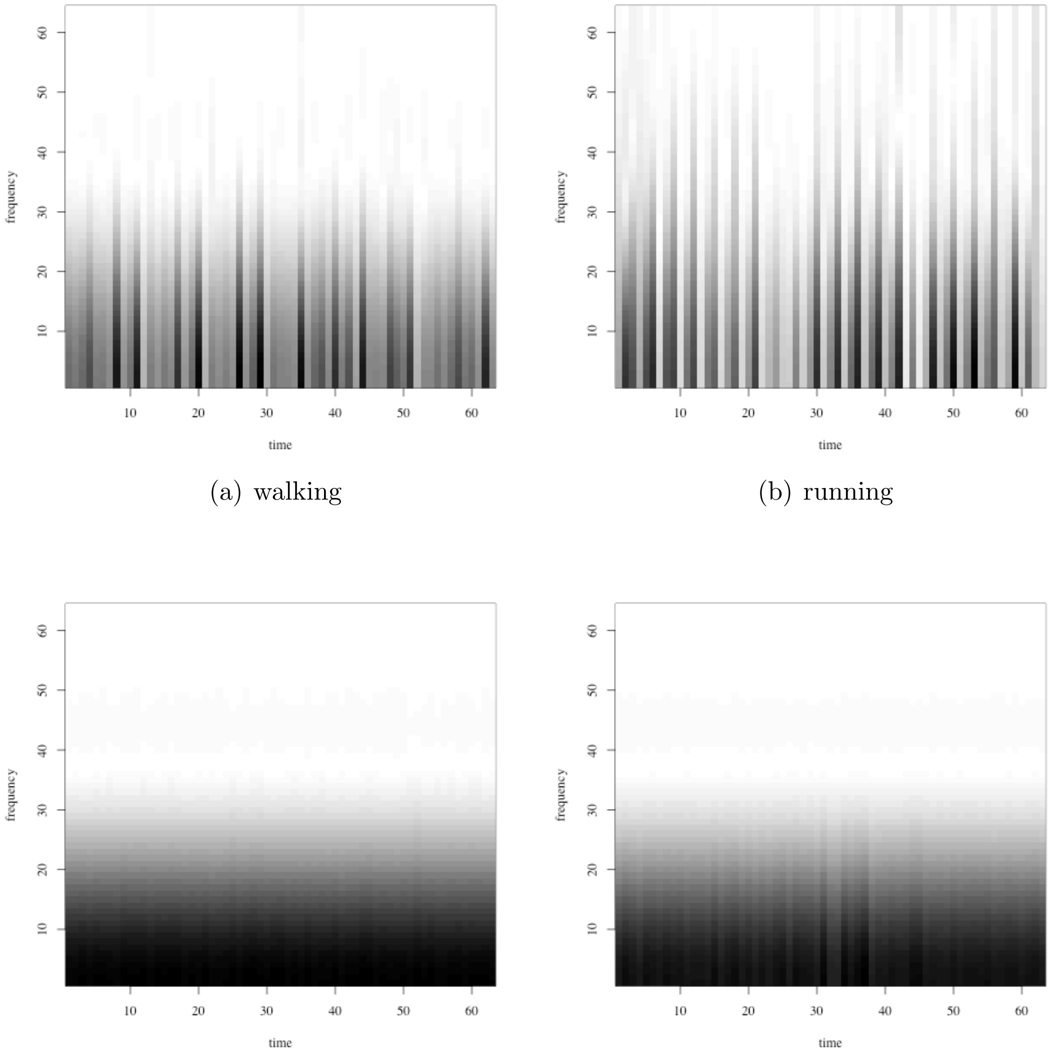
An alternative of the wavelet transform is the short-time Fourier transform (STFT). We visualized them to compare as below.

Fig. 2.8 shows examples of the STFT for the same signals as Fig. 2.4. The difference between "standing still" and "in a moving train" in Fig. 2.8 as that in Fig. 2.4. This is because the time duration of the STFT should be fixed over the whole time course. This means that the STFT has a low time resolution for low frequency and a high time resolution for high frequency, which is not suitable to extract the difference between the confusing states. In fact, the accuracy of STFT is 25%, which is much lower than ours.

2.5 Conclusion

We proposed a method for mobile devices that recognize the user's contexts. This requires only a single three-axis accelerometer and extracts features using the wavelet transform with the Haar mother wavelet. The feature space is reduced to two dimensions by the SVD, where the largest and the second largest singular values are features. The two features are classified by a multi-layer perceptron to one of the four user-states, walking, running, standing still and being in a moving train.

We investigated the dependency of its performance and computation time on the sampling rate and the time-window width by experiments with public data. The proposed method discriminated the user's contexts with accuracy of over 90% in most cases, which is better than the conventional method although its computation load is comparable.



(a) walking

(b) running

(c) standing still

(d) being in a moving train

Figure 2.8. Example of STFT. The x- and y-axis represent the time and the frequency.

Chapter 3

Environment Recognition on Mobile Devices

3.1 Introduction

In our daily life, we are exposed to many types of sounds including environmental sound, which has both positive and negative impact on our health although the relationship of noise to the human environment is complex. Health is defined by the Constitution of the World Health Organization as ‘a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity’ [19], suggesting that sound exposure affects not only our health but also quality of life and well-being. In many studies, sound levels are used as noise metrics, which represents loudness of the sound. However, they are just one factor of sounds and are not enough for understanding the relationship between environmental sound and our health. In order to analyze the relationship from several points of view, we capture the sound in real environment with mobile phones, because mobile devices, including mobile phones and smart phones, have become indispensable in our daily lives. People carry their mobile devices almost everywhere at all times. One of the advantages of using the environmental sounds is that we are able to address the following technical problems simultaneously: a) to recognize environment wherever people go to; and b) to identify the health status of the environment. Other sensors such as Global Positioning System (GPS) or sound-level meter can not address the problems at the same time. In the case of GPS, we can only know the environment but can not identify health status. Sound-level meter can be useful to identify the health status but not to know the environment. In addition, some mobile phones are not equipped with some advanced mobile sensors. On the other hand, all mobile phones are equipped with a microphone sensor, which allows to capture environment sounds. Although many research addresses either environmental sound recognition or sound exposure and health, we consider both of them for mHealth application.

To explore health understanding of environmental sounds, we derive sound health analysis as three correlated sound pattern recognition: a) routine and environment classification, to recognize which environment that the sound is recorded from; b) space categorization, to categorize open or closed space of the sound; and c) health distinction, to distinguish sound health impact in terms of people opinion. In this chapter, we propose using transfer learning for health understanding. Transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve correlated new problems faster or with better solutions. The core of transfer learning is the modeling of knowledge transfer (KT), which aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task. We expect that a system can smartly address the above three tasks at the same time, then the system must have a good knowledge of environmental sound toward health understanding.

3.2 Related Work

3.2.1 Sound Exposure and Health

Noise exposure could induce hearing impairment, hypertension and ischemic heart disease, annoyance, sleep disturbance, and decreased school performance [20]. Table 3.1 gives an overview of the noise effects research based on the report by the Committee on Noise and Health, an international committee of the Health Council of the Netherlands [21]. A scientific evidence of causal relationship between noise and health has been rated in terms of ‘sufficient’, ‘limited’, ‘inadequate’, or ‘lack’, respectively. The observation threshold as the lowest noise exposure level also has been defined only if the evidence is sufficient.

Table 3.1. Long term effects of noise exposure

Effect	Scientific Evidence	Measurement/ Threshold	Reference
Hearing impairment	Sufficient	$L_{Aeq,24h}$, 70	[22]
Hypertension	Sufficient	L_{den}, 70	[21]
Ischaemic heart disease	Sufficient	L_{den}, 70	[21]
Biochemical effects	Limited	N/A	[21]
Immune effects	Limited	N/A	[21]
Birth weight	Limited	N/A	[21]
Prenatal disorders	Lack	N/A	[21]
Annoyance	Sufficient	L_{den}, 42	[23]
Psychosocial well-being	Limited	N/A	[21]
Performance at School	Sufficient	$L_{Aeq,school}$, 70	[20]

Noise can be described by various metrics such as the A-weighted decibel scale, sound level equivalents, day-night average sound levels, and percentile levels. Since the human hearing organ is not equally sensitive to sounds of different frequencies, the sound pressure level (L) is 'A-weighted' and expressed as dB(A), which is the most common metrics of sound and environmental noise. The A-weighted equivalent continuous sound level is denoted L_{Aeq}. If the level is normalized to an 8-hour workday, it is denoted $L_{Aeq,8h}$. If it is over a period of T hours, then it is denoted $L_{Aeq,T}$ and is defined as follows:

$$L_{Aeq,T} = 10 \log \left(\frac{1}{T} \int 10^{L(t)/10} \right) dt \tag{3.1}$$

where $L(t)$ is the A-weighted sound level at time t and T is the duration of the exposed period in seconds. The day-evening-night equivalent level is denoted L_{den} and is defined as follows:

$$L_{den} = 10 \log \frac{1}{24} \left(12 * 10^{L_{day}/10} + 4 * 10^{(L_{evening}+5)/10} + 8 * 10^{(L_{night}+10)/10} \right) \tag{3.2}$$

where L_{day} is the A-weighted average sound level over the 12 hour day period of 7 – 19 h, $L_{evening}$ is the A-weighted average sound level over the 4 hour day period of 19 – 23 h, and L_{night} is the A-weighted average sound level over the 8 hour day period of 23 – 7 h. $L_{evening}$ and L_{night} have an adjustment of 5 and 10 dB(A) respectively to take account of the difference in annoyance due to the time of day.

As seen above, noise impacts our health in the ways as listed in Table 3.1. However, no deterministic relationship has been discovered so far about how environmental sound leads to a certain health status. This paper discovers feature and conducts classification towards predictive environmental sound health understanding. In addition, we propose using transfer learning for health understanding.

3.2.2 Environmental Sound Recognition

Environmental sound recognition has received more attention in recent years. Several features have been used to describe audio signals. Mel-frequency cepstral coefficients (MFCCs) are one of the popular features in audio classification and speech recognition. Due to a lack of a standard database for environmental sound recognition, MFCCs are often used by researchers for benchmarking their work [24]. Other commonly-used features for audio signals include Zero-crossing rate (ZCR), Short-time average energy, Spectral centroid, Bandwidth, Band-energy ratio, Spectral roll-off, Linear prediction coefficients (LPC), and Cepstral coefficients [25].

Peltonen et al. [25] classified auditory scenes into 17 scenes out of 26 scenes, which were Street, Road, Nature, Construction site, Market place and Amusement park for outdoors, and Car, Bus, Train and Subway train for vehicles, and Restaurant/cafe, Pub, Supermarket, Lecture pause and Crowd/indoors for public/social places, and Office, Lecture/meeting and Library for offices/meeting rooms/quiet places, and Living room, Kitchen, Bathroom and Music for home, and Church, Railway Station, Subway station and Hall for reverberant. For data collection, they considered various configurations: a binaural setup (a Brel & Kjaer 4128 head and torso simulator), a stereo setup (AKG C460B microphones), and a B-format setup (Sound-Field MkV microphone). They recorded the sounds on a digital multitask recorder with a 16-bit, 48-kHz sampling rate and a Sony (TCD-D10) digital audio tape recorder with a 16-bit, 48-kHz sampling rate. They made a comparison with different features such as LPC, Band-energy and ZCR, and showed that MFCCs used in conjunction with Gaussian Mixture Model (GMM)-based classifier performed well for an auditory scene recognition experiment involving identifying 17 different auditory scenes from 26 scenes. They obtained a recognition accuracy of 63.4%. Ma et al. [26] classified the acoustic environment into 10 environments, which were Office, Lecture, Bus, Urban driving, Railway station, Beach, Bar, Laundrette, Soccer match and City center street. A high quality microphone and portable recorder were used to capture the audio examples. They used MFCC features and a hidden Markov model (HMM) classifier, and achieved over 90% accuracy although humans averaged only 35% on the same data. Chu et al. [27] proposed to use the Matching Pursuit (MP) based algorithm to obtain effective time-frequency features. They compared the recognition accuracy using MP, MFCC and their combination for 14 classes of sounds, which were Inside restaurants, Playground, Street with traffic and pedestrians, Train passing, Inside moving vehicles, Inside casinos, Street with police car siren, Street with ambulance siren, Nature-daytime, Nature-nighttime, Ocean waves, Running water/stream/river, Raining/shower, and Thundering. They obtained sound clips from BBC Sound Effects Library - Original Series [28] and the Freesound Project [29]. For classification, k-nearest neighbor and GMM classifiers were tested. The MP features performed better than MFCC. By combining MP and MFCC features, they obtained a recognition accuracy of 83.9%. They also compared with other commonly used features such as ZCR, Band-energy ratio, and Spectral centroid. The average recognition accuracy was 55.2%, which was much worse than using combined MFCC and MP features. In a subsequent paper by Chu et al. [30], they proposed a framework for a composite of deep belief networks (composite-DBNs) to recognize 12 different types of everyday environments obtained sound clips from BBC Sound Effects Library - Original Series [28], which were Inside casino, Playground, nature-daytime, Inside restaurants, Next to rivers/streams, Train passing, Inside vehicles, Raining, Street with traffic, Ocean waves, and Thundering. MP-features and MFCC were used, and DBN was compared with GMM. The DBN provided the better performance.

Our target application is on life environment sound and health analysis. Despite all previous efforts on environmental sound recognition, sound is recorded by special device. This does not satisfy the scenario of our life environment application. Moreover, there is no previous work reviewed in this section on predicting sound relation to health. On the other hand, our work is environmental sound health understanding, which covers firstly environment recognition, and then sound and health correlation analysis.

3.3 Environmental Sound Health Understanding

For health understanding, we firstly consider the basic problem, that is, which environment the sound comes from. As we know, some environment such as park and beach have mostly positive impact on our health, whereas others like airport noise have negative impact. Thus environment, in some sense, determines the health impact of sound. To address this, we specifically observe routine including going to the office/school, relaxing in a park, and doing some exercise, as sound may generate because of some specific human (i.e., people who conduct sound recording) activities. We also observe environment (i.e., the location of the sound) including street, park, and shopping mall, as sound comes from objects in a specific environment.

On top of environment and routine, we also look at the space characteristic of the sound. Sound from open space differs to that from closed space, in terms of not only sound spectrum but also health impact. At this point, we observe characteristic of each environment as indoor, outdoor, nature, and transport. Here we separate natural environment from outdoor, because nature does not include the sound related to technology. Nature is almost equivalent to the label of health. And we specifically look at environment related to transport because transport is one of the important daily life activities, and it includes both indoor and outdoor.

Most importantly, we certainly consider health distinction. In practice, it is often difficult for us to judge the health impact of one type of sound, unless we conduct a long term medical experiments on people. For the convenience, we simply distinguish sound health impact in terms of the comfortableness to people, as comfortableness in most cases represents the health, according to [19]. For the convenience of expression, we will use the term "health" instead of comfortableness throughout the rest of paper.

Although the environment and space characteristic does not determine the health status of sound directly, the above three concepts are correlated in a way that environment and space characteristic of sound present different types of evidence to health status; and compared to the environment concept, the space characteristic of sound gives more sense of health status.

It is natural to think about using knowledge of environment and space characteristic to facilitate the health understanding of environmental sound, in terms of machine learning. Thus, we consider three classification tasks, environmental classification, space categorization, and health distinction, and conduct transfer learning for an enhanced health understanding.

Specifically, this includes three transfer learning as (1) environment to health; (2) space to health; and (3) environment to space, where the third transfer learning guarantees the validity of the first two knowledge transfer. Fig. 3.1 gives the general system dialog of transfer learning based environmental sound health understanding.

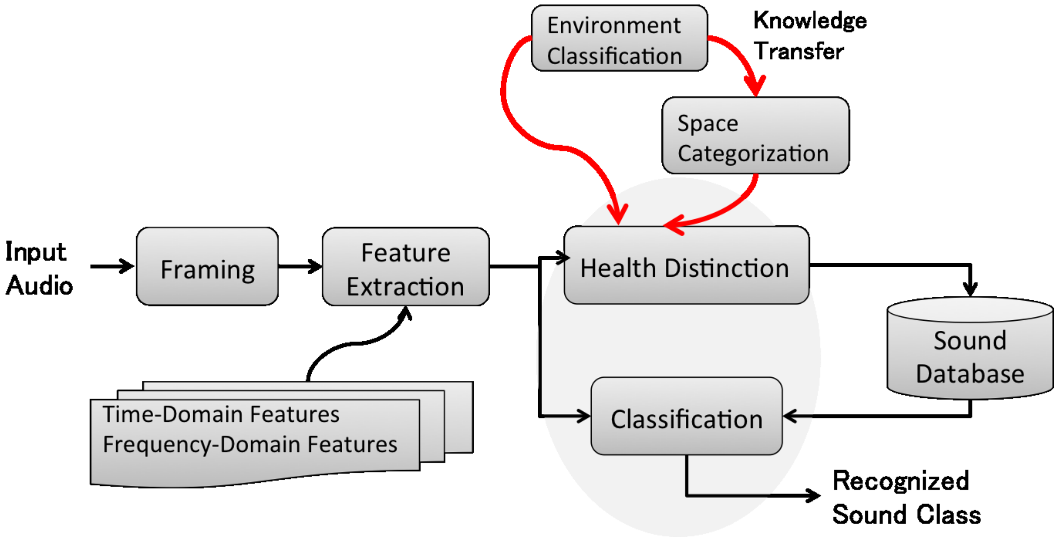


Figure 3.1. System dialog of transfer learning based environmental sound health understanding

3.4 Audio Feature Extraction for Environmental Sound Classification

Signals have both time and frequency domain representations.

3.4.1 Time-Domain Features

Root Mean Square (RMS) [31, 32], often used as a measure of loudness, is defined as follows:

$$RMS = \sqrt{\frac{\sum_{n=1}^{N} F_n^2}{N}} \tag{3.3}$$

where N is the number of samples in a frame, and F_n is the value of the n-th sample of a frame. It is computationally inexpensive and easy to implement.

Zero-crossing rate (ZCR), an indicator for the noisiness of the signal, is defined as the number of times that a signal changes signs within a particular frame, which has been widely used in voice activity detection, voiced/unvoiced speech classification and music/speech classification [33, 34]. It also, together with RMS, can be used to make a simple speech/no speech distinction. ZCR is calculated as follows:

$$ZCR = \frac{1}{N} \sum_{n=1}^{N} |sgn[F_n] - sgn[F_{n-1}]| \tag{3.4}$$

where sgn[.] is a signum function, N and Fn are as defined previously.

3.4.2 Frequency-Domain Features

Frequency-Domain features are calculated using the frequency spectrum of a signal.

Spectral Roll-off point [32, 35], a measure of the skewness of the spectral sharp, is defined as the frequency R_t below which a certain amount (threshold) of the magnitude distribution is concentrated

$$\sum_{n=1}^{R_t} M_t[n] = P * \sum_{n=1}^{N} M_t[n] \tag{3.5}$$

where $M_t[n]$ is the magnitude of Fourier transform at frame t and frequency bin n, P is the threshold in percentage, and N is the number of samples in a frame. The threshold in our experiments is 0.85.

Spectral Centroid (SC) [32, 35], a measure of brightness and general spectral shape, represents the balancing point of the spectral power distribution. The SC for a frame is computed as follows:

$$SC = \frac{\sum_{n=1}^{N} n * |M_t[n]|^2}{\sum_{n=1}^{N} |M_t[n]|^2} \tag{3.6}$$

where N, $M_t[n]$ and n are as defined previously.

Bandwidth (BW) [36] is computed as the magnitude weighted average of the distance between the SC and the spectral components. BW is defined as

$$BW = \frac{\sum_{n=1}^{N} (n - SC)^2 * |M_t[n]|^2}{\sum_{n=1}^{N} |M_t[n]|^2} \tag{3.7}$$

where N, $M_t[n]$, n and SC are as defined previously.

Band Energy Ratio (BER) [37] is the ratio of the energy in a certain frequency-band to the total energy. BER is defined as

$$BER = \frac{\sum_{n=K_1}^{n=K_2} |M_t[n]|^2}{\sum_{n=1}^{N/2-1} |M_t[n]|^2} \tag{3.8}$$

where K_1 and K_2 are the frequency points of the given frame, N and $M_t[n]$ are as defined previously. In our experiments, four logarithmic sub-bands are used.

MFCCs [38] were extracted applying the discrete cosine transform (DCT) to the log-energy outputs of mel-scaling filter-bank.

3.5 Transfer Learning towards Sound Health understanding

Transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve correlated new problems faster or with better solutions. The core of transfer learning is the modeling of knowledge transfer (KT), which aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task. Note that the roles of the source and target tasks are not symmetric, transfer learning in contrast to multitask learning cares most about the target task [39].

3.5.1 Existing Knowledge Transfer Approaches

Research on transfer learning has attracted more and more attention since 1995 [40]. Based on characteristics of KT bridges, the different approaches in the literature can be divided into the following categories: inductive bias sharing approaches, memory item sharing approaches, and probability sharing approaches.

In inductive bias sharing approaches, the learning system gives a prior assumption to the previous knowledge, which is considered as an inductive bias to knowledge transfer implementation [41]. In memory item sharing approaches, the learning system performs knowledge transfer based on the training examples stored in long-term memory, known as memory items [42]. In probability sharing approaches, the learning system utilizes the hierarchical Bayesian framework to provide knowledge transfer for a new learning task [43]. These knowledge transfer approaches mentioned above attempt to discover the relatedness between tasks into an embedding learner/classifier for MTL. This type of method is called a learner-dependent knowledge transfer model [44].

The problem with the previous knowledge transfer methods is that the process of transferred knowledge is not transparent. Various classifiers have the advantages in addressing different data distributions, and no single classifier can perform well in all classification problems. Thus from the viewpoint of learner independence, transferred knowledge is irrelevant to the learner, which is considered to be essential for a desirable knowledge transfer.

3.5.2 Raw Data Knowledge Transfer

With the assumption that two correlated tasks shares the same feature space, Pang et al. [44] proposed a new KT bridge, where the knowledge in transfer is decomposed as raw data, thus can be incorporated into any learner as additional training data input to facilitate the learning rate.

3.5.3 Task Relatedness Calculation

Let T^0 be a primary task, and T^k be a secondary task with training data $D^0 = [X^0, Y^0]$, and $D^k = [X^k, Y^k]$, respectively. Theoretically, $k = 1, \ldots, m$ as there certainly exist more than one task correlated to T^0. In KT research, $k = 1$ as a total of two tasks are given for KT. The relatedness R^{0k} of T^0 and T^k is typically defined over the available training samples and the hypotheses for these related tasks as,

$$R^{0k} = f_R(\mathcal{L}(D^0), \mathcal{L}(D^k), D^0, D^k), \tag{3.9}$$

where f_R can be either a static relatedness measure such as Hamming Distance or Linear Coefficient of Correlation, or a dynamic measure, between the developing hypothesis of the primary task and that of the secondary task. $\mathcal{L}$ is a learning system for MTPR, which could be any type of classifier, e.g., in ηMTL [45], it is specified as an ANN.

In raw data KT, task relatedness is independent to any classifiers/learners, thus we exclude the influence of $\mathcal{L}$ in (3.9) as,

$$R^{0k} = f_R(D^0, D^k). \tag{3.10}$$

The correlation of tasks is defined as the set of samples that are mutually beneficial to perform the learning task. Specifically, given subspace S^0 spanned by a subset of D^0, cast S^0 into T^k space, if S^0 in T^k space, denoted as $S^{0\to k}$ has no 'class confliction' for T^k, then S^0 is correlated to T^k, and the correlation $C^{0\to k}$ is extracted by S^0 as,

$$\begin{gathered} C^{0\to k}(S^0) = \arg\max_{S^0 \in S_{D^0}} |S^0| \\ \forall(\vec{x}^k, \vec{y}^k) \in S^{0\to k}, y^k \equiv c, \end{gathered} \tag{3.11}$$

where $|S^0|$ represents the size of S^0, and S_{D^0} is a space spanned by D^0. 'class confliction' here is interpreted as $\forall(\vec{x}^k, \vec{y}^k) \in S^{0\to k}, y^k \equiv c$, in which $(x^k, y^k) \in [X^k, Y^k]$, and c is a class label from T^k.

Summarizing all subsets related to T^k in D^0, we have the correlation of T^0 to T^k as

$$\mathcal{C}^{0\to k} = \bigcup_{\forall S^0 \in S_{D^0}} C^{0\to k}(S^0), \tag{3.12}$$

which is reflected as a complete set correlation from T^0 to T^k. Because of the symmetry between the primary and secondary tasks, we can also have $\mathcal{C}^{k\to 0}$ from the above definition.

3.5.4 Knowledge Carrier

In practice, the above raw data KT relies on knowledge (i.e., selected raw data) carrier/container (KC), while knowledge transferring.

Given dataset D^0 and D^k from two correlated tasks T^0 and T^k respectively, for any subset $d^0 \subset D^0$ in one class, a subspace is spanned as,

$$B^0_{c,r} = KC(d^0_i) \tag{3.13}$$

where c is the center of the space and r is the radius, by which the $B^0_{c,r}$ is able to tell whether a new input instance is enclosed by the KC or not.

To verify the utility of $B^0_{c,r}$ for T^k, we cast the KC into T^k data space, and we have

$$B^{0\to k}_{c^{0\to k},r^{0\to k}} = CAST(B^0_{c,r}, D^0, D^k) \tag{3.14}$$

where $B^{0\to k}_{c^{0\to k},r^{0\to k}}$ is the resulting KC from casting $B^0_{c,r}$ in T^k data space, and the CAST function is implemented by calculating the casting KC center $c^{0\to k}$ and the casting KC radius $r^{0\to k}$, respectively.

$$c^{0\to k} = (c^0 - c^k \frac{r^k_{\max}}{r^0_{\max}}), \tag{3.15}$$

and

$$r^{0\to k} = \frac{r^k_{\max}}{r^0_{\max}} r^0. \tag{3.16}$$

where $r^0_{\max}$ is the radius of KC over D^0, and $r^k_{\max}$ is the radius of KC over D^k.

The obtained $B^{0\to k}_{c,r}$ is expected to cast a subset S^k instances in D^k. $B^0_{c,r}$ is judged as a sharable data space by T^k, if all instances of S^k belong to one class in T^k. The instances enclosed by $B^0_{c,r}$ are the correlation data of T^0 to T^k. In this way, given $\forall d^0 \subset D^0$ the entire sharable data is obtained as a merge of all KCs that satisfy the correlation definition and the smoothness assumption [46] as: given two instances located in a high-density region, if one is enclosed in a sharable KC, so for the other instance,

$$\begin{aligned} B_x^* = \{b_i^0\} \cup \{x\} \ \textit{subject to} \ b_i^0 \in \text{one of } D^1\text{class, and } b_i^{0\to k} \in \text{one of } D^k\text{class} \\ d(c, x_j) > r, d(c, x_i) < r, \text{and } d(x_i, x_j) < \theta \end{aligned} \tag{3.17}$$

where θ is a distance threshold that represents the density of data distribution.

3.5.5 Environmental Sound Health Analysis

We derive sound health analysis as three correlated sound pattern recognitions: (1) routine and environment classification, to recognize which environment that the sound is recorded from; (2) space categorization, to categorize open or closed space of the sound; and (3) health distinction, to distinguish sound health impact in terms of people opinion.

Let T^E, T^O and T^H be the above three tasks respectively. We set up on purpose three transfer learning $T^E \to T^H$, $T^O \to T^H$ and $T^E \to T^O$, in finding those set of data that have always positive contribution to the health understanding. Note that although $T^E \to T^O$ is not health targeted transfer, the positive transfer of $T^E \to T^O$ ensures the validity of the rest of two transfers towards health understanding.

In doing that, we apply eq (3.12) to the above three transfers respectively, and collect data sharable across all three tasks as,

$$\begin{aligned} \mathcal{C} &= \mathcal{C}^{E\to H} \cup \mathcal{C}^{O\to H} \\ &\quad \textit{subject to } \mathcal{C}^{E\to O} \textit{is positive.} \end{aligned} \tag{3.18}$$

With $\mathcal{C}$, systems are able to maximize health understanding of environmental sound, based on a good environment and space recognition rate. In our system implementation, we used Minimum Enclosing Ball [47, 48] as the knowledge carrier.

3.6 Mobile Environment Sound Data Collection

Environmental sound in daily life is our interest since we are exposed to many types of sounds in our lives, such as the sounds from TV, household appliances, and traffic.

The use of portable electronic devices such as mobile phones, tablet computers and personal digital assistants (PDAs) has been widely recognized as an efficient way to improve the provision of healthcare. These mobile devices are equipped with powerful embedded sensors, such as accelerometers, Global Positioning System (GPS), microphones, and cameras, which are enabling new applications across a wide range of domains, including business, healthcare, social networks, safety, environmental monitoring, and transportation [49].

3.6.1 Setup

Consider mobile phone is widely used in our life and actually it has become an important part of our lives, we conveniently choose smart phone as voice recording device. To ensure good location coverage, our data collection is designed to be based on people's daily life routine, which enables to cover a wide range of life environments including public transport, office, street, and shopping mall. In addition, the timepoint of sound recording is another important factor that we care. As we know, environmental sound varies over time in real world practice. For example of swimming pool environment, the number of swimmers in public holidays is often several times than that in working days, which produces rather different levels of sound. Thus for each environment/routine, several recording sessions are conducted at different times to ensure the time coverage of data collection.

We collected environmental sound data mainly in Auckland, New Zealand during the Autumn of 2014. The devices that we used for sound recording are two types of smart phone which include iPhone 5 (sampling rate 44.1 kHz) and Samsung Galaxy S4 mini (sampling rate 16 kHz). For each observed environment, we conducted multiple sessions of sound recording at different locations as well as different time and/or dates. For example of park environment, we recorded sound in the morning and afternoon for 7 different parks of Auckland. In general, addressing 17 routines of 13 environments, we conducted total 121 sessions of recording at 43 different locations in Auckland. The obtained sound data forms our experimental dataset whose statistical information is given in Tables 3.2 and 3.3.

3.6.2 Labelling Data in Multi-task

To explore health understanding of environmental sounds, we labeled obtained data by manually skimming the sound. Firstly, we labeled them with the environment which the sound belongs to, including train, beach, shopping mall, and street, as shown in Table 3.2.

Then, we defined on purpose four bigger categories: transport, outdoor, indoor and nature, which is the space characteristic of the sound (Table 3.3). Sound from open space differs to that from closed space, in terms of not only sound spectrum but also health impact. Here we separate natural environment from outdoor, because nature does not include the sound related to technology. Nature is almost equivalent to the label of health. And we specifically look at environment related to transport because transport is one of the important daily life activities, and it includes both indoor and outdoor. The space characteristic is close to our target concept although it is not equivalent.

Finally, we labeled data with positive (P) or negative (N) impact on our health, according to the feeling of people who did the sound recording (Table 3.4). Because in practice, it is often difficult for us to judge the health impact of one type of sound, unless we conduct a long term medical experiments on people.

We expect that a system can smartly address the above three tasks at the same time, then the system must have a good knowledge of environmental sound toward health understanding.

3.7 Experiments

3.7.1 Setup

For each experiment, we use a 10-fold cross-validation, where accuracy is averaged over 10 runs and at each run, one tenth of the data is used as a testing set and the rest as the training set. The analysis window length for all features was 1024 ms and the used windowing function was Hamming window. The overlap between successive frames was 50% of the frame length. k-Nearest Neighbors (kNN) are the conceptually simplest of classifiers [50]. kNN are a simple algorithm that uses the majority vote of the k nearest training patterns to assign a class label.

Table 3.2. Routine and Environment

No	Routine	Environment	Duration (minute)
1	Train to school/office	Train	376
2	Walk in beach	Beach	38
3	Drive to school/office	Car	8
4	Walk near MW	Street	34
5	Bus to school/office	Bus	180
6	Wait for train	Train Station	6
7	Bus in city	Bus	100
8	Shopping in mall	Shopping mall	14
9	Restaurant	Restaurant	4
10	Walk to school/office	Street	116
11	Walk in park	Park	218
12	Trip	Plane	2
13	Relax in park	Park	254
14	Swimming pool	Pool	220
15	Walk in city	Street	60
16	Walk in library	Library	16
17	Exercise	Gym	2
			1648 in total

Table 3.3. Health Related Category

Health Related Category	Environment	Duration (minute)
Transport	Plane	2
	Train	376
	Car	8
	Bus	280
Outdoor	Street	210
Indoor	Train Station	6
	Shopping mall	14
	Restaurant	4
	Swimming pool	220
	Library	16
	Gym	2
Nature	Beach	38
	Park	472

The proposed system is implemented on the platform of Matlab R2012a, and the experiments are carried out on a PC with an Intel Core i7 1.7GH_Z CPU and 8G-byte memory.

3.7.2 Optimal number of Sound Duration Time

The recognition rates are obtained for the 17 environments using different features with the kNN classifier. For kNN, we used the Euclidean distance as the distance measure and the 17-nearest neighbor queries to obtain the results. As features, ZCR, RMS, SR, SC, BW, BER(4) and MFCC(20) are used. The F-measures under 5-fold cross validation are shown in Table 3.5. The training and test sequence duration were 1 to 30 seconds, respectively. We obtained a recognition accuracy of 87.08% with 20 MFCC features. With all of the features, we obtained a recognition accuracy of 88.19%. From this result, 3 seconds are enough for classification.

Table 3.4. Health and Environment

People's opinion	Environment	Duration (minute)
N	Train	330
	Beach	4
	Car	0
	Street	210
	Bus	242
	Train Station	6
	Shopping Mall	14
	Restaurant	4
	Park	60
	Plane	2
	Swimming Pool	220
	Library	0
	Gym	2
P	Train	46
	Beach	34
	Car	8
	Street	0
	Bus	38
	Train Station	0
	Shopping Mall	0
	Restaurant	0
	Park	412
	Plane	0
	Swimming Pool	0
	Library	16
	Gym	0

Table 3.5. Average F-measure with different features using the kNN

	ZCR	RMS	SR	SC	BW	BER(4)	MFCC(20)	ALL
1 sec	0.1400	0.1570	0.1615	0.1497	0.1463	0.3662	0.8529	0.8637
2 sec	0.1496	0.1604	0.1680	0.1518	0.1469	**0.3699**	0.8703	0.8769
3 sec	0.1525	**0.1718**	0.1700	0.1561	0.1467	0.3651	**0.8708**	**0.8819**
5 sec	0.1576	0.1659	0.1696	0.1557	0.1516	0.3687	0.8641	0.8688
10 sec	0.1643	0.1713	**0.1798**	**0.1724**	0.1530	0.3339	0.8295	0.8362
20 sec	**0.1756**	0.1703	0.1737	0.1601	0.1564	0.3133	0.7199	0.7368
30 sec	0.1657	0.1510	0.1773	0.1596	**0.1585**	0.3152	0.6391	0.6264

3.7.3 Optimal number of MFCCs

The optimal number of the MFCC coefficients was examined for each task. Fig. 3.2 shows the example of MFCCs range from 1 to 300. Since the latter half of the plot has the same tendency as the first half, we only analyze the number of MFCCs up to 180. We calculated the F-measures with several number of MFCCs. Note that MFCC 1 to 5 were used in the case of MFCC(5), and similarly MFCC 1 to 100 were used in the case of MFCC(100), for instance.

Fig. 3.3 shows the accuracy of each number of MFCCs for three tasks: (a) Impact on Health, (b) Environment, and (c) Indoor, Outdoor, Transport, Nature. As seen, using MFCC(15), MFCC(25) and MFCC(20) gives better accuracy for each task.

3.7.4 Importance of Fundamental Features

As mentioned in Section 3.4, ZCR, RMS, SR, SC, BW and BER(4) are often used to capture the characteristics of environmental sound in addition to MFCCs. We call them fundamental features in this paper. In order to confirm the importance of fundamental features, we compared the accuracy with three conditions: only fundamental features, only MFCCs, and both fundamental features and MFCCs.

Fig. 3.3 shows the result of the comparison for each task. Although only using the fundamental features shows poor accuracy, the combination of fundamental

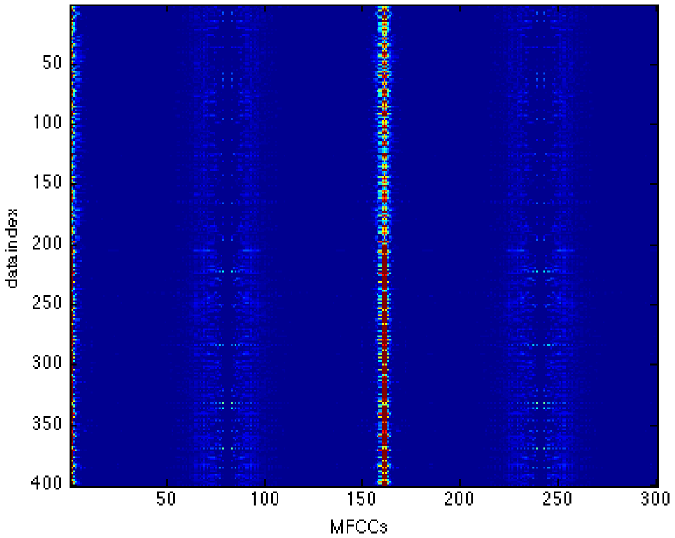


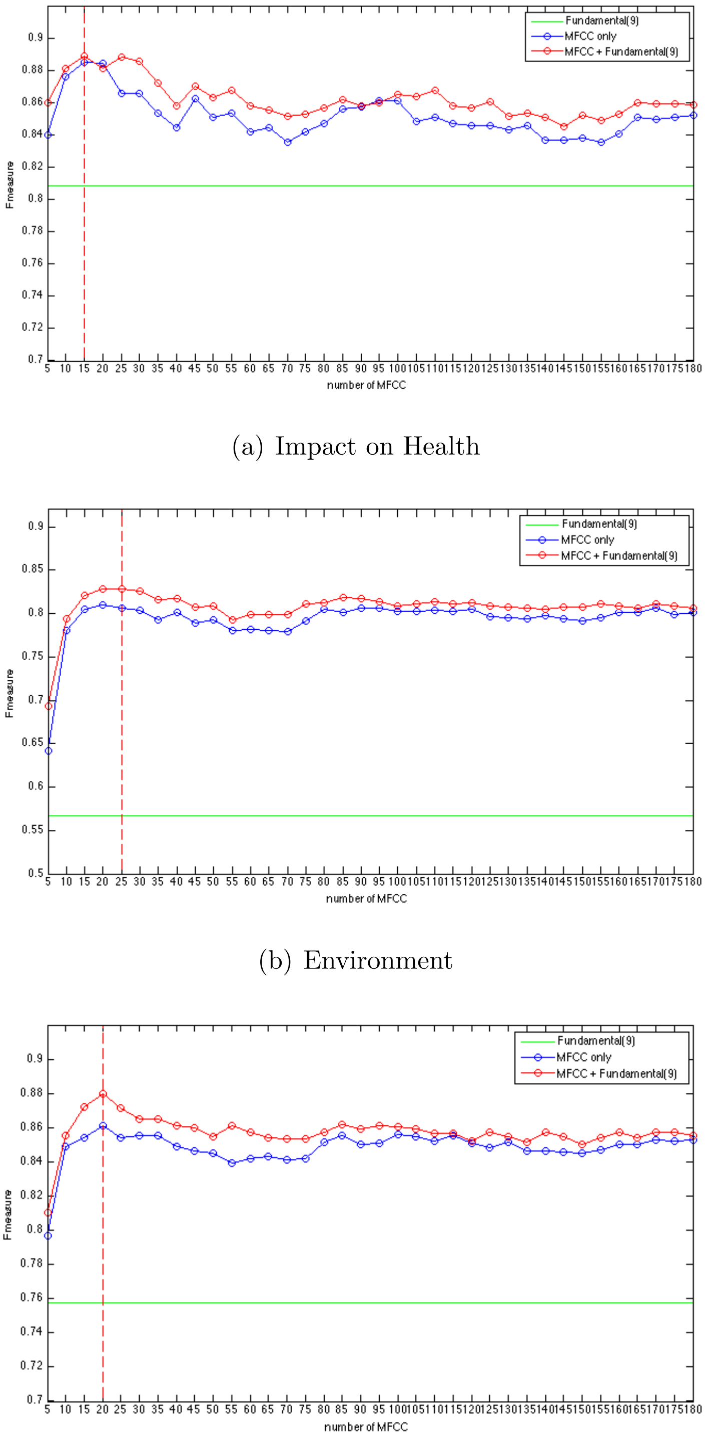
Figure 3.2. Example of MFCCs range from 1 to 300

features and MFCCs gives better accuracy than only MFCCs, which implies fundamental features are also important.

Fig. 3.4 shows the more details. We compared with and without MFCCs against each fundamental feature, fundamental features (fnd(9)). Based on the result of previous experiment, MFCC(15), MFCC(25) and MFCC(20) are used for each task, respectively. The red dashed line in Fig. 3.4 represents the accuracy of using only MFCCs, and the '+' after the feature name means with MFCCs. As seen, although each fundamental feature shows less accuracy, the combination of fundamental features and MFCCs gives better accuracy than only MFCCs, which implies again that fundamental feature set gives some contribution.

3.7.5 Knowledge Transfer

For each cross validation, we define two classification tasks. We set one task as the primary task, and the other task is set as the secondary task. Knowledge transfer (KT) is always conducted from the primary task to the secondary task. The obtained correlated data is then used as additional training data for the

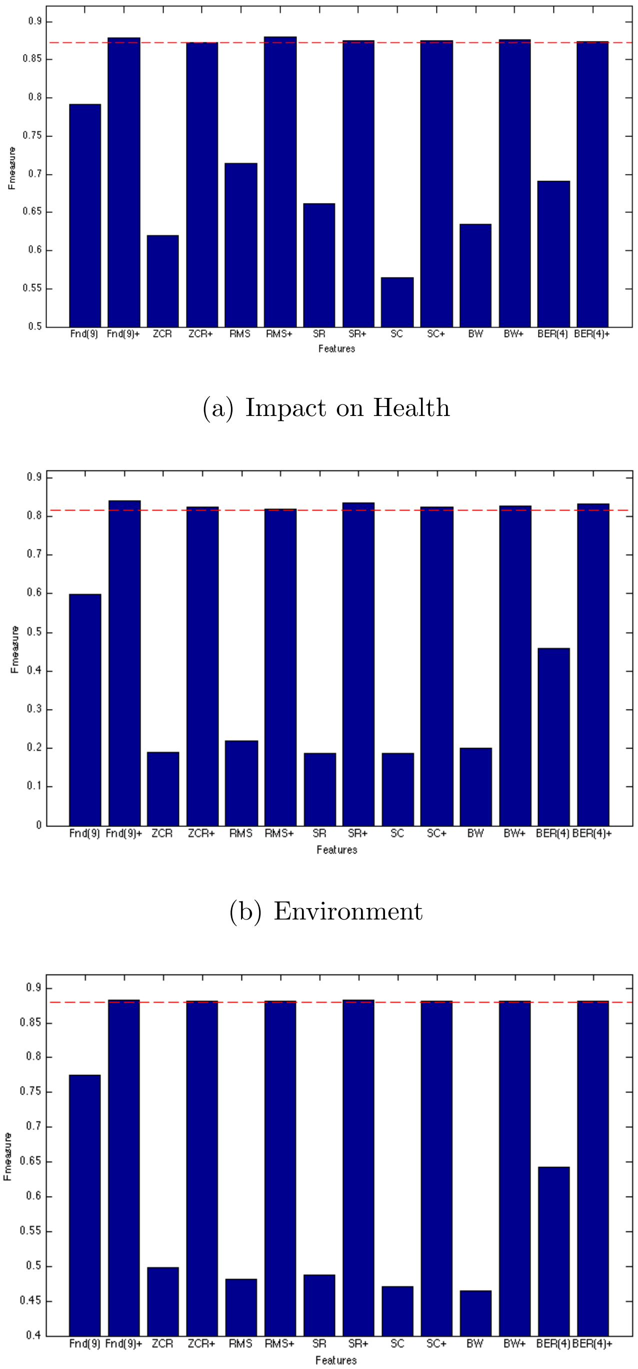


(a) Impact on Health

(b) Environment

(c) Indoor, Outdoor, Transport, Nature

Figure 3.3. Accuracy Comparison for each task



(a) Impact on Health

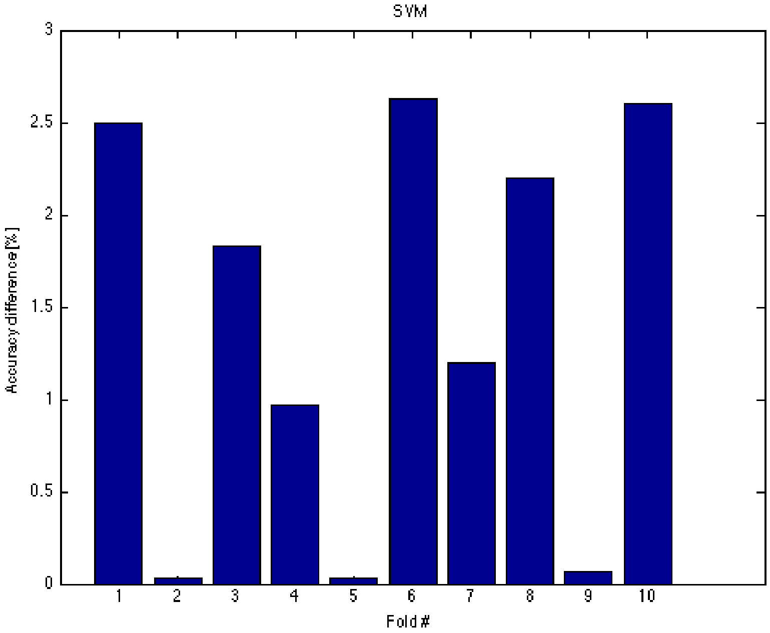
(b) Environment

(c) Indoor, Outdoor, Transport, Nature

Figure 3.4. Contribution of Fundamental Features

secondary task. As the proposed KT is classifier independent, classifiers with different characteristics are applied to MTL. For comparison, we report the results of MTL without KT. The proposed MEB-based KT algorithm is implemented on the platform of Matlab 7.80 (R2009a), and the experiments are carried out on a PC with an Intel(R) Core(TM) i7-2600 3.40GH_Z CPU and 8G-byte memory.

We did three transfer learning as (1) environment to health; (2) space to health; and (3) environment to space, where the third transfer learning guarantees the validity of the first two knowledge transfer. Fig. 3.5 shows the classification accuracy differences between MTLs with and without KT. As seen in Fig. 3.5, positive transfer is happened on SVM. This indicates that through transfer learning, health understanding of environmental sound is enhanced.



(a) Environment to health

Figure 3.5. Accuracy difference

3.8 Conclusion

This chapter addressed a new mobile Health (mHealth) application that helps people track the daily activity, and determines the health status automatically for each activity people experienced. Instead of relying on advanced mobile sensors, we focused on a microphone with which all mobile phones are equipped, and collected the environmental sounds toward health understanding. In order to discover new knowledge in sound feature space and improve health understanding of our daily life environment, we perform a knowledge transfer (KT), the key of multi-task learning. The experimental results show that KT with knowledge carrier perform better than without KT.

Future work includes the implementation on real mHealth application.

Chapter 4

Conclusion

We considered an mHealth application toward helping people track their daily activities. In order to build such applications, in this disseration, we solved the following technical problems: a) recognition of human activity and b) recognition of the user's environment. We first proposed a new approach to recognize user activity for mobile applications. The key of the proposed method is using a three-axis accelerometer, which is equipped with almost any mobile devices. In order to keep high accuracy in recognition with low computational cost, We employed the wavelet transform and the singular value decomposition during feature extraction. We investigated the dependency of its performance and computation time on the sampling rate and the time-window width by experiments with public data. Our method discriminated the user activity with accuracy of over 90% in most cases, which is better than the conventional method although its computation load is comparable. We also discussed the mother wavelet and an alternative of the wavelet transform, which is the STFT. Our experimental results showed that the calculation time of Haar mother wavelet is much shorter than others although the accuracy is almost the same. As for the STFT, the accuracy using STFT is 25%, which is much lower than the accuracy using our proposed method. These results indicate that the proposed method can be successfully applied to commonly used mobile devices.

We also proposed a new approach to recognize the user's environment for mobile applications. Instead of relying on advanced mobile sensors, we focused on a microphone with which all mobile phones are equipped, and collected the environmental sounds. To ensure sufficient location coverage, our data collection was designed based on people's daily routines, which enable coverage of a wide range of environments including public transportation, offices, streets, and shopping malls. In order to classify the 17 environments, we employed several audio features from time domain and frequency domain, and obtained over 80% accuracy for the 17 environments.

Future work includes the implementation on real mHealth application. When dealing with real applications, it is necessary to remove or reduce the noise during the operation of mobile devices. One of the possible solutions is to filter out the noise. However, we need to take into account energy consumption.

In addition to environment recognition, environmental sound in daily life towards health understanding is our interest since we are exposed to many types of sounds in our lives, such as the sounds from TV, household appliances, and traffic. In order to improve health understanding of our daily environment, we need to conduct furthur analysis of environmental sound.

Acknowledgements

This dissertation could not have been completed without the great support of many people. First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Kazushi Ikeda, who offered me invaluable assistance, continuous encouragement and patience. I am deeply grateful to Prof. Shoji Kasahara and Prof. Keiichi Yasumoto for reviewing this dissertation and providing feedback. I would like to thank Prof. Hossein Sarrafzadeh and Prof. Paul Pang for accepting me as a visiting scholar at Unitec. Studying in New Zealand is a valuable experience for me. I would like to express my appreciation to Associate Professor Yuichi Sakumura, Dr. Takatomi Kubo and Dr. Tomoya Tamei, who helped me with their useful comments to improve my dissertation. I am grateful to Ms. Aya Tanimoto for taking care of most of the administrative things allowing me to concentrate on research. My special thanks go to Prof. Hideo Konami, who encouraged and motivated me to begin this journey. Furthermore, I thank all the members of the Mathematical Informatics Laboratory and Decentralized Machine Learning Intelligence Laboratory for helpful discussion and suggestion. Finally, I would like to thank my family, host family and friends, especially Mizuki and Inah, for their love, concern and support.

Bibliography

[1] "White paper on information and communications in japan," Ministry of Internal Affairs and Communications, Tech. Rep., 2013.

[2] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," vol. 48, no. 9, pp. 140–150, 2010.

[3] J. F. M. Bernal, L. Ardito, M. Morisio, and P. Falcarin, "Towards an efficient context-aware system: Problems and suggestions to reduce energy consumption in mobile devices," in *Proc. ICMB-GMR 2010*, 2010, pp. 510–514.

[4] E. J. Y. Wei and A. T. S. Chan, "Towards context-awareness in ubiquitous computing," in *Proc. EUC 2007*, 2007, pp. 706–717.

[5] T. Iso and K. Yamazaki, "Gait analyzer based on a cell phone with a single three-axis accelerometer." Proc. MobileHCI 2006, 2006, pp. 141–144.

[6] K. Cho, N. Iketani, H. Setoguchi, and M. Hattori, "Human activity recognizer for mobile devices with multiple sensors." Proc. ATC 2009, 2009, pp. 114–119.

[7] G. Dedes and A. G. Dempster, "Indoor GPS positioning challenges and opportunities." Proc. IEEE VTC 2005, 2005, pp. 412–415.

[8] Healthkit. Accessed January 29, 2015. [Online]. Available: https://developer.apple.com/healthkit/

[9] Google fit. Accessed January 29, 2015. [Online]. Available: https://developers.google.com/fit/

[10] P. Goupillaud, A. Grossman, and J. Morlet, "Cycle-octave and related transforms in seismic signal analysis," *Geoexploration*, vol. 23, no. 1, pp. 85–102, 1984.

[11] J. Mantyjarvi, J. Himberg, and T. Seppanen, "Recognizing human motion with multiple acceleration sensors." Proc. IEEE SMC 2001, 2001, pp. 747–752.

[12] Y. Washizawa, “Feature extraction using constrained approximation and suppression,” *IEEE Transactions on Neural Networks*, vol. 21, no. 2, pp. 201–210, 2010.

[13] M. Sun, J. Yang, C. Liu, and J. Yang, “Similarity preserving principal curve: An optimal 1-D feature extractor for data representation,” *IEEE Transactions on Neural Networks*, vol. 21, no. 9, pp. 1445–1456, 2010.

[14] D. E. R. et al., *Parallel Distributed Processing.* MIT Press, 1987.

[15] D. M. Hawkins, “The problem of overfitting,” *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, 2004.

[16] D. F. Shanno, “Conditioning of quasi-Newton methods for function minimization,” *Math. Comput.*, vol. 24, no. 111, pp. 647–656, 1970.

[17] G. E. Moore, “Cramming more components onto integrated circuits,” *Electronics*, vol. 38, no. 8, p. 114117, 1965.

[18] P. L. Inc. iphone, ipad, and ipod benchmarks. [Online]. Available: http://browser.primatelabs.com/ios-benchmarks

[19] “Official records of the world health organization,” the Constitution of the World Health Organization, Tech. Rep., 1946, accessed April 17, 2014. [Online]. Available: http://www.who.int/about/definition/en/print.html

[20] W. Passchier-Vermeer and W. F. Passchier, “Noise exposure and public health,” *Environmental Health Perspectives*, vol. 108 (suppl 1), no. 4, pp. 123–131, 2000.

[21] “Noise and health,” The Hague: Health Council of the Netherlands, Health Council of the Netherlands: Committee Noise and Health, Tech. Rep., 1994.

[22] “Acoustics - determination of occupational noise exposure and estimation of noise-induced hearing impairment,” 1990.

[23] M. HME and O. CGM, “Annoyance from transportation noise: Relationships with exposure metrics dnl and denl and their confidence intervals,” *Environmental Health Perspectives*, vol. 109, no. 4, pp. 409–416, 2001.

[24] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," in *Proceedings of 2013 Annual Summit and Conference on Asia Pacific Signal and Information Processing Association (APSIPA 2013)*, 2013, pp. 1–9.

[25] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002)*, vol. 2, 2002, pp. 1941–1944.

[26] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Trans. on Speech and Language Processing*, vol. 3, no. 2, pp. 1–22, 2006.

[27] S. Chu, S. S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[28] The bbc sound effects library original series. Accessed June 18, 2014. [Online]. Available: http://www.sound-ideas.com/bbc.html

[29] The freesound project. Accessed June 18, 2014. [Online]. Available: http://www.freesound.org/

[30] S. Chu, S. Narayanan, and C. C. J. Kuo, "Composite-dbn for recognition of environmental contexts," in *Proceedings of 2012 Annual Summit and Conference on Asia Pacific Signal and Information Processing Association (APSIPA 2012)*, 2012, pp. 1–4.

[31] J. Harrington and S. Cassidy, *Techniques in Speech Acoustics.* Kluwer, 1999.

[32] M. Cord and P. Cunningham, *Machine learning techniques for multimedia: case studies on organization and retrieval.* Springer, 2008.

[33] J. D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analysis for musical instrument classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 38, no. 2, pp. 429–438, 2008.

[34] P. Kathirvel, M. Manikandan, S. Senthilkumar, and S. Soman, "Noise robust zerocrossing rate computation for audio signal classification," in *Proceedings of 3rd International Conference on Trendz in Information Sciences and Computing (TISC)*, 2011, pp. 65–69.

[35] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[36] A. Ramalingam and S. Krishnan, "Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 4, pp. 457–463, 2006.

[37] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[38] S. Molau, M. Pitz, R. Schlter, R. S. Uter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2001)*, 2001, pp. 73–76.

[39] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[40] Nips1995 workshop. learning to learn: Knowledge consolidation and transfer in inductive systems. Accessed September 15, 2014. [Online]. Available: http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html

[41] T. Mitchell, "The need for biases in learning generalizations," Department of Computer Science, Rutgers University, Tech. Rep. CBM-TR-117, 1980.

[42] S. Ozawa, A. Roy, and D. Roussinov, "A multitask learning model for online pattern recognition," *IEEE Trans. on Neural Networks*, vol. 20, no. 3, pp. 430–445, 2009.

[43] K. Yu, A. Schwaighofer, V. Tresp, W.-Y. Ma, and H. Zhang, "Collaborative ensemble learning: Combining collaborative and content-based information filtering via hierarchical bayes," in *Uncertainty in Artificial Intelligence: Proceedings of the 19th Conference (UAI-2003).* Morgan Kaufmann, 2003, pp. 616–623.

[44] S. Pang, F. Liu, Y. Kadobayashi, T. Ban, and D. Inoue, "A learner-independent knowledge transfer approach to multi-task learning," *Cognitive Computation*, November 2013.

[45] D. L. Silver, "Selective functional transfer: Inductive bias from related tasks," in *IASTED International Conference on Artificial Intelligence and Soft Computing (ASC2001.* ACTA Press, 2001, pp. 182–189.

[46] O. Chapelle, A. Zien, and B. Scholkopf, Eds., *Semi-Supervised Learning.* MIT Press, 2006.

[47] M. Badoiu and K. L. Clarkson, "Optimal core-sets for balls," in *Proc. 14th ACM-SIAM Sympos. Discrete Algorithms*, 2003, pp. 801–802.

[48] P. Kumar, J. S. B. Mitchell, and E. A. Yildirim, "Approximate minimum enclosing balls in high dimensions using core-sets," *ACM Journal of Experimental Algorithmics*, vol. 8, 2003.

[49] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell, "A survey of mobile phone sensing," *Communications Magazine, IEEE*, vol. 48, no. 9, pp. 140–150, 2010.

[50] A. Sogaard, *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2013.

Publication List

Journal Papers

1. Yuki Maruno, Kenta Cho, and Kazushi Ikeda. Energy-Efficient User-State Recognition Method Using Wavelet Transform and Singular Value. SICE Journal of Control, Measurement, and System Integration (SICE JCMSI), vol.8, no.1, pp. 86–92, 2015.

International Conferences and Workshops

1. Yuki Maruno, Kenta Cho, Yuzo Okamoto, Hisao Setoguchi, and Kazushi Ikeda. Human Activity recognition Using Wavelet Transform and Singular Value Decomposition. ITC-CSCC2011, Gyeongju, Korea, June 2011.

2. Yuki Maruno, Kenta Cho, Yuzo Okamoto, Hisao Setoguchi, and Kazushi Ikeda. An Online Human Activity Recognizer for Mobile Phones with Accelerometer. ICONIP2011, Shanghai, China, November 2011.

Domestic Conferences and Workshops

1. 丸野由希, 池田和司, Paul Pang. ユーザの生活環境モニタリングに向けた環境推定手法. 計測自動制御学会 システム・情報部門学術講演会 2014 (SSI2014), 岡山, 2014年11月.