SpiderWalk: Circumstance-aware Transportation Activity Detection Using A Novel Contact Vibration Sensor

LIANG WANG, State Key Laboratory for Novel Software Technology, Nanjing University, China
WEN CHENG and LIJIA PAN, School of Electronic Science and Engineering, Nanjing University, China
TAO GU, School of Computer Science and Information Technology, RMIT University, Australia
TIANHENG WU, XIANPING TAO, and JIAN LU, State Key Laboratory for Novel Software Technology, Nanjing University, China

This paper presents the design and implementation of the *SpiderWalk* system for circumstance-aware transportation activity detection using a novel contact vibration sensor. Different from existing systems that only report the type of activity, our system detects not only the activity but also its circumstances (e.g., road surface, vehicle, and shoe types) to provide better support for applications such as activity logging, location tracking, and smart persuasive applications. Inspired by but different from existing audio-based context detection approaches using microphones, the *SpiderWalk* system is designed and implemented using an ultra-sensitive, flexible contact vibration sensor which mimics the spiders' sensory slit organs. By sensing vibration patterns from the soles of shoes, the system can accurately detect transportation activities with rich circumstance information while resisting undesirable external signals from other sources or speech that may cause the data assignment and privacy preserving issues. Moreover, our system is implemented by reusing existing audio devices and can be used by an unmodified smartphone, making it ready for large-scale deployments. Finally, a novel temporal and spatial correlated classification approach is proposed to accurately detect the complex combinations of transportation activities and circumstances based on the output of each individual classifiers. Experiments conducted on a real-world data set suggest our system can accurately detect different transportation activities and their circumstances with an average detection accuracy of 93.8% with resource overheads comparable to existing audio- and GPS-based systems.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**;

Additional Key Words and Phrases: Crack-resistance Sensor, Circumstance-aware, Transportation

ACM Reference Format:
Liang Wang, Wen Cheng, Lijia Pan, Tao Gu, Tianheng Wu, Xianping Tao, and Jian Lu. 2018. SpiderWalk: Circumstance-aware Transportation Activity Detection Using A Novel Contact Vibration Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 42 (March 2018), 30 pages. https://doi.org/10.1145/3191774

1 INTRODUCTION

Detecting users' transportation activities has been an emerging field which can support many applications including activity and energy expenditure estimation [2], crowd mobility analysis and prediction [12, 15], persuasive applications [1, 8], and etc. While existing work on transportation activity/mode detection mainly focuses

Authors' addresses: Liang Wang, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China, wl@nju.edu.cn; Wen Cheng; Lijia Pan, School of Electronic Science and Engineering, Nanjing University, Nanjing, China, ljpan@nju.edu.cn; Tao Gu, School of Computer Science and Information Technology, RMIT University, Australia, tao.gu@rmit.edu.au; Tianheng Wu; Xianping Tao; Jian Lu, State Key Laboratory for Novel Software Technology, Nanjing University, China, wth@smail.nju.edu.cn,{txp,lj}@nju.edu.cn.

2474-9567/2018/3-ART42 $$15.00
https://doi.org/10.1145/3191774

(a) Pedestrians are using a new mud road not planned by the government to cross the lawn

(b) People walk on a brick 'bridge' over a tarmacadam road after a heavy rain

Fig. 1. Circumstance information of transportation activities (road surface in this case) can help the government users to improve their services. For (a), a brick pathway is needed to help people walking on the mud road; For (b), timely traffic control and road management are required. (Pictures from the Internet.)

on the type of activity alone, this paper concerns the problem of detecting users' transportation activities with rich circumstance information (e.g., road surface, shoe, and vehicle types). Our work is motivated by the observation that even the routine transportation activities may have varying circumstances which provide important information for both government and personal users. Government users such as the transportation management departments can receive detailed and timely reports on transportation-related events by tracking peoples' transportation activities and circumstances to provide better services such as road management and traffic control. For example, Fig. 1 illustrates two examples of how the circumstances of transportation activities can reveal important information to improve government services which cannot be obtained with traditional techniques. Human investigation may provide the same information but is slow and costly. For personal users, circumstance-aware transportation activity detection can enable new applications with greatly improved performance and user experience. For an example of walking activity, walking in a room wearing dress shoes (*pacing*) and walking in the woods wearing boots (*hiking*) may imply different energy expenditure models, and should be treated differently. In addition, knowing what surface a user is walking on, e.g., a tarmacadam road or a brick sideway, is critical to applications such as pedestrian safety monitoring [16]. In another example, by knowing that one is wearing high heels, a smart persuasive application may suggest the user to rest constantly to prevent potential injuries. In our daily work or physical exercise, the risks of injuries caused by the environment and the footwear have been recognized by health experts for a long time [18, 36]. However, very few persuasive or coaching applications incorporate such knowledge due to the lack of information.

To facilitate such smart applications, the underlying detection system should provide not only the type of activity (e.g., walking, running or idle) but also its circumstances such as shoe, road surface, and vehicle types. Existing work mainly adopt sensors such as GPS/GIS [11, 31, 41], accelerometer [15, 30, 32, 38], barometer [29], and magnetic-field sensors [6] to detect the gross body movements. However, this approach fails to discover detailed circumstances because such knowledge is out of the reach of these sensors' sensing abilities. Insole sensing approaches that use force [33] and capacitive sensors [22] have the potential of performing activity and floor type sensing. However, they are still limited in their detection scopes as discussed later in Sec. 2.

To fill this niche, in this paper we propose *SpiderWalk*, a system which is built using a novel contact vibration sensor for circumstance-aware transportation activity detection. The proposed sensor is built by tailoring a novel crack-resistance sensing material which mimics the sensory slit organs located at the leg joints of spiders [17]. Manufactured by depositing a thin layer of platinum (*Pt*) with deliberately formed cracks on top of a polyurethane acrylate (PUA) substrate, the crack-resistance sensing material is ultra-sensitive in capturing subtle vibration signals. In [17], the authors show the novel sensing material's potential in supporting a wide-range of applications including sensing the strings' vibrations of musical instruments, capturing human speech when

attached to the neck, and sensing the pulses by attaching to the wrist. By deploying such a sensor under a user's foot with high-frequency sampling (i.e., 8kHz), our system can detect the subtle vibration patterns caused by different activities under different circumstances including vehicle, road surface, and shoe types using audio-based context sensing approaches [21, 24]. To the best of our knowledge, our work is among the first to perform such detailed activity and circumstance detection using vibration sensing.

To build a system with high performance while facing the many challenges brought by the real-world applications, we carefully design and implement *SpiderWalk* with many design considerations as discussed in Sec. 3. First, to accurately capture vibration signals for various transportation activities under different circumstances, we propose to attach the sensor under the foot based on the experience that the feet are sensitive to vibration differences under different transportation modes. We validate the effectiveness of this design by real-world experiments and comparing with a microphone-based sensing system built following the state-of-the-art audio-based approaches [21, 24]. Moreover, we address the data assignment and privacy preserving issues of traditional audio-based systems through this design by only detecting vibrations from the sole of a shoe. Because our vibration sensor is attached to the foot inside the shoe, it is resistant to external sounds produced by other sources around the subject of interest (SoI) and the user's speech that propagates through the air. Further, we pay special attention to the user comfort issue because our system is worn under the feet during daily lives. State-of-the-art contact microphones built with stethoscope augmented microphones [39] or brass piezoelectric sensors [26] are uncomfortable to wear because they are rigid and large in size. Different from these systems, the crack-resistance sensing material adopted in our system is thin (less than 1mm) and flexible which can be attached to the foot like a bandage as shown in Sec. 5. Finally, we address the device simplicity issue by hacking a commercial-off-the-shelf (COTS) Bluetooth audio adapter to perform vibration sampling, encoding, and transmission without requiring specialized hardware support as did in existing contact microphone designs [26]. This is done by tailoring the crack-resistance sensing material into a vibration sensor that matches the output standards of a COTS electret microphone used in various audio devices. Following this design, the *SpiderWalk* system is implemented by simply replacing the electret unit of a COTS Bluetooth audio adapter with our sensor and reusing the existing audio sampling, encoding, transmission, and processing devices, making it low-cost and ready-to-use in real-world applications.

Different from existing work on transportation activity detection which only predicts the activity itself, our work aims at addressing the more complex problem of deciding the combination of transportation activities and various circumstances. To address this issue, we propose to combine the results of four independent classifiers through a novel integration framework that explores the temporal and spatial correlations among different classes to obtain a unified detection result. Experiment results suggest the proposed approach can recover the classification errors made by the independent classifiers and significantly improve the detection accuracy.

We conduct extensive experiments on real-world data, and the results suggest that *SpiderWalk* performs accurate detection of various transportation activities and their circumstances with an average detection accuracy of 93.8%. The resource overheads of the system are comparable to existing audio sensing systems [26] and GPS positioning modules on smartphones [20], suggesting that *SpiderWalk* provides richer information without increasing the system's resource consumption compared to existing audio- and GPS-based systems. To reduce the system's overheads on CPU, memory and power consumption, we propose the frame admission control and feature selection techniques which can be integrated into the data processing pipeline. We also discuss several possible ways to further reduce the system's overheads. As presented in Sec. 7, many potential applications can be drawn from this system including but not limited to map generation and tracking, personal assistance, security and health applications.

In summary, this paper makes the following contributions.

- Our main contribution is the design and implementation of a novel contact vibration sensor using a ultra-sensitive crack-resistance sensing material and COTS devices;
- We benchmark the performance of the proposed sensor and compare it with a COTS electret microphone;
- We implement the *SpiderWalk* system for circumstance-aware transportation activity detection based on this novel sensor;
- A temporal and spatial correlated classification approach is proposed to accurately detect complex combinations of transportation activities and their circumstances;
- Extensive experiments are conducted on real-world data to evaluate the system's performance.

The rest of the paper is organized as follows. Sec. 2 introduces the related work. Sec. 3 presents our motivation and considerations when designing the system. Detailed system design and implementation are introduced in Sec. 4 and Sec. 5, respectively. Sec. 6 presents our experiment results. We discuss the potential applications in Sec. 7. Finally, Sec. 8 concludes the paper.

2 RELATED WORK

We summarize the related work in this section. We categorize existing work into two classes: 1) work on transportation activity and mode detection that explores different sensing modalities to achieve goals related to our work; 2) audio-based context and activity sensing approaches that use microphones to capture audio data which are close related to vibrations used in this work.

2.1 Transportation Activity/Mode Detection

Much research work has been conducted for transportation activity and mode detection using different sensors. We summarize the existing work into four categories: 1) location-based; 2) smartphone sensor-based; 3) magnetic-field sensor-based; and 4) insole sensing-based approaches.

2.1.1 Location-based. In [11], Gong et al. present a GPS-based travel survey study combined with GIS information in New York City that identifies users' activities including *walk*, *subway*, *rail*, *car*, and *bus*. Zheng et al. [41] propose to use motion related features extracted from GPS data to detect *walk*, *bike*, *bus*, and *driving* activities. Shah et al. use GPS data combined with acceleration for motorized transportation mode identification which includes *car*, *bus*, and *train* [31]. In [27], Reddy et al. present a transportation mode classification system that also combines GPS and acceleration data. Their system can accurately detect transportation modes including *still*, *walk*, *run*, *bike*, and *motor*.

Different from GPS-based approaches, some researchers propose to use the location information provided by the cellular network for location-based transportation mode detection [3, 35]. In [35], the coarse-grained GSM traces are used to detect general movements such as *stationary*, *walking*, and *driving*. In [3], the authors use information collected from the cellular towers to recognize activities including *stationary*, *car*, *train*, and *walk*.

While existing work has shown their effectiveness, location-based approaches are limited in several aspects. First, they rely on the availability of GPS or cellular signals and cannot perform accurate estimation in indoor or remote places with poor signal strength. Second, as discussed in the introduction, location-based approaches are only capable of detecting the gross movements of the subjects. Fine-grained circumstance information such as road surfaces and shoe types are beyond their sensing abilities.

2.1.2 Smartphone Sensor-based. To increase the system's availability and reduce its reliance on infrastructures, recent work has explored smartphones' built-in sensors for transportation mode detection.

Hemminki et al. [15] introduce a hierarchical classification system to discriminate modes including *stationary*, *walk*, *bus*, *train*, *metro*, and *tram* using acceleration data. In [30], Shafique et al. propose to use smartphones' acceleration data for transportation mode (*walk*, *bicycle*, *car*, and *train*) prediction. To further reduce power

consumption, Sankaran et al. propose to use smartphones' barometers to detect simple transportation modes including *idle*, *walking* and *vehicle* [29].

The advantages of acceleration- and barometer-based approaches lie in their low power demand and can provide continuous detection service without infrastructure support. However, similar to the location-based approaches, these approaches are limited in detecting fine-grained circumstance information mainly due to the low sensitivity and sampling rates of the sensors. As shown later in Sec. 6.6.3, experiment results suggest an inertial-based approach can perform accurate detection of transportation modes including different activities and vehicle types. However, it suffers from low detection accuracies for other circumstance information such as shoe and road surface types.

2.1.3 Magnetic-field Sensor-based. In [6], Chen et al. propose the Mago system that can perform reliable transportation mode detection by fusing the metal-induced magnetic field distortion captured by the Hall-effect magnetic-field sensors and the vibrations captured by the smartphones' built-in accelerometers. Their approach can accurately detect seven classes of commute activities (*stationary*, *bus*, *bike*, *car*, *train*, *light rail* and *scooter*). Moreover, their system can differentiate the phone's in-car position which is an important piece of context information. While SpiderWalk shares a similar data processing pipeline with Mago, our sensing approaches are fundamentally different. Our approach can detect transport related circumstance information even if it does not have an observable effect on the magnetic-field (e.g., shoes).

2.1.4 Insole Sensing. In [33], Shu et al. propose an in-shoe pressure sensing array to detect actions such as *normal standing*, *standing on one leg*, *heel strike*, and *push off*. Their work shows the potential of in-shoe pressure measurements for transportation activity detection. Combining insole sensors with smartphones' built-in sensors, Zhang et al. [40] fuse readings from a novel foot force sensor with smartphone's GPS data for fine-grained transportation mode recognition. Their system can discriminate transportation modes including *walking*, *cycling*, *bus passenger*, *car passenger*, and *car driver*.

Different from forces, insole capacitive sensors have also shown their effectiveness in sensing the foot movements [34], which makes them potentially useful for transportation activity detection. Because the capacitive sensor readings are influence by the type of floors, skin conductivity, and properties of the socks [22, 34], Matthies et al. [22] propose the CapSoles system that can detect the user identities and floor types with capacitive pressure-sensitive insoles. Experiment results suggest CapSoles can perform accurate detection of floor types including *sand*, *lawn*, *pavement*, *tartan*, *linoleum*, and *carpet*. The system can also detect postures including *standing*, *sitting*, *kneeling*, *lying*, and *carrying*.

Our system is similar to the above insole sensing approaches in two aspects. First, we also use an insole sensing solution to perform activity and floor type sensing, and share the advantages including user comfort, accuracy, and infrastructure independence [22, 33]. Second, our work is close to CapSoles [22] in that we are both interested in detecting the floor types. Different from the above work that uses force and capacitive sensors, in this work, we propose to use an ultra-sensitive vibration sensor with a high sampling rate to capture transportation related circumstance information. As a result, our system is expected to be more sensitive in detecting the subtle differences in various transportation activities, and provide richer circumstance information.

In summary, the major difference of our work from existing work lies in that we use a novel insole contact vibration sensor to detect transportation activities with rich circumstance information which is beyond the scope of existing work.

2.2 Audio-based Context & Activity Sensing

Another category of work closely related to this work is audio-based context and activity sensing. Researchers from the Dartmouth College have conducted a series of work on audio-based context sensing on smartphones

using the built-in microphones including SoundSense [21], DSP.Ear [9], and DeepEar [19]. Their work shows that audio-based context sensing can be performed on smartphones with high performance and efficiency. Instead of using smartphones' resources alone, Auditeur [24] is proposed to provide a mobile cloud service for power-efficient acoustic event detection on smartphones.

Different from the above work that uses smartphones' built-in microphones, recent work has explored various contact microphones for activity and context sensing. BodyScope is proposed in [39] that uses stethoscope augmented microphones to detect different activities from non-speech body sounds. In [26], Rahman et al. introduce the design and implementation of a novel contact microphone built on top of a brass piezoelectric sensor to capture various non-speech body sounds for activity and context recognition.

Motivated by and different from the above work, we propose in this work the design and implementation of a novel contact vibration sensor built using a novel, ultra-sensitive crack-resistance sensing material [17] to capture different vibration patterns caused by various transportation activities under different circumstances.

3 MOTIVATION & CONSIDERATIONS

In this section, we present the motivation and design considerations of our system, and briefly introduce the way our system meets these considerations. Detailed system design and implementation are introduced later in Sec. 4 and Sec. 5, respectively.

3.1 Motivation

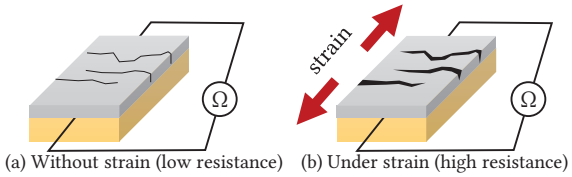
Motivated by the recent developments in audio-based sensing technologies which have shown to be effective in recognizing various ambient contexts and human activities [21, 26, 39], our basic idea is to apply audio sensing technologies for circumstances-aware transportation activity detection. Common physics tells us that sound is produced by the vibration of the audio source which propagates by air pressure variation. People can easily tell the difference between walking wearing high-heels and athletic shoes by simply listening to the sounds of steps. Moreover, such vibration is propagated through not only the air but also medias that make contact with the vibration source. For example, distinctive vibration patterns can be felt by the feet when sitting in the car and the metro, caused by the engine and the wheels running on railways, respectively.

Motivated by the above work and life experiences, in this work, we propose to use a foot-worn contact vibration sensor to capture the vibration patterns caused by different transportation activities under different circumstances. While the idea behind this work is simple, there are many challenges in designing and implementing a practical wearable system that can achieve our goal, which we discuss in detail through our design considerations as follows.

3.2 Data Assignment & Privacy Preservation

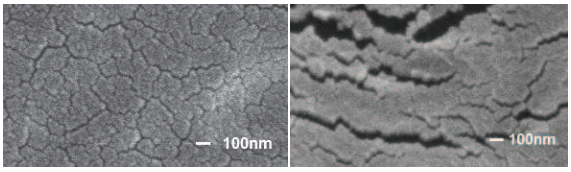
In our context, the data assignment issue is expressed as whether the sensed data are actually produced by the subject of interest (SoI). It is challenging for a microphone to accurately assign the data because it captures not only the sounds produced by the SoI but also sounds produced by other sources around the SoI. For example, an audio-based system may mistakenly determine a user sitting in a busy lobby to be walking on capturing the stepping sounds of other pedestrians around the user.

Another challenge for a common microphone is the privacy preservation issue. While a microphone is mainly designed and optimized to capture human speech for recording or communication purposes, it becomes a major threat to privacy when speech is not intended to be captured. Though audio-based context sensing systems [21] are often built for capturing ambient sounds, it is often inevitable that the sensitive conversation data are also recorded by the microphones.



(a) Without strain (low resistance) (b) Under strain (high resistance)

Fig. 2. Crack-resistance sensing material, grey layer above—*Pt* layer, yellow layer below—PUA layer, black lines on *Pt* layer—cracks.



(a) Without strain (low resistance) (b) Under strain (high resistance)

Fig. 3. Scanning electron microscope image.

The root-cause of a common microphone to suffer from the data assignment and privacy preservation issues is its sensitivity to external sound/noise. And it is difficult to address the issues because its nature of capturing sounds by sensing the ambient air pressure variations. In [26], the authors show the advantage of a piezoelectric-based contact microphone in blocking external sounds and noise, making their system potentially applicable in our scenario. However, their solution is also limited as discussed next.

3.3 User Comfort & Device Simplicity

User comfort is an important issue when designing a wearable system. While contact microphones outperform common microphones on resisting external sounds, they are uncomfortable to wear and complex in system implementation. For example, stethoscope augmented microphones [39] are large in size and impossible to wear under the foot. In [26], the authors developed a piezoelectric sensor based contact microphone which are small in size and sensitive to sounds propagated through flesh and muscle. However, the brass piezoelectric sensor used in their work is rigid and large in size. Film piezoelectric sensors are good candidates to build wearable systems for being thin and flexible. However, piezoelectric sensor based systems suffer from the device simplicity issue as discussed next.

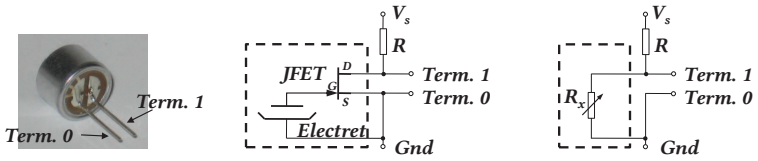
Prototype systems built using experimental, customized devices are often less optimized, higher in cost, and less robust compared to COTS devices. As a result, it is desirable to use COTS devices to build a system to make it ready-to-use in large-scale, real-world applications. Like audio-based context sensing systems, an easy way to implement a system is to use the existing audio sensing devices and processing frameworks to collect and process vibration data. Existing piezoelectric-based sensors, however, cannot fit into a common audio device for there is a mismatch between the output voltage of the sensors and common electret microphones as we will explain later in Sec. 4.1.2. As a result, fully customized systems are built using specialized sampling and processing hardware [26], making them complex and costly. Different from piezoelectric sensor-based approaches, the contact vibration sensor used in our system is designed to simulate a COTS electret microphone. The proposed system can then be implemented by simply replacing the built-in microphone of a common audio device with our sensor, and reusing the remaining parts, making it low-cost, and ready-to-use for real-world applications.

4 SENSOR DESIGN & EVALUATION

Following the above design considerations, this section introduces our contact vibration sensor design and performance benchmarking results.

4.1 Contact Vibration Sensor Design

Our contact vibration sensor is built by tailoring a novel crack-resistance sensing material to simulate a common electret microphone.



(a) Appearance of an electret microphone

(b) Simplified schematic with an electret unit

(c) Equivalent circuit of an electret microphone

Fig. 4. Electret mic. vs. variable resistor.

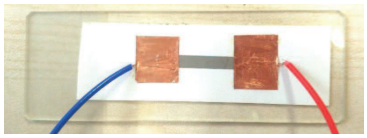


Fig. 5. Contact vibration sensor.

4.1.1 Crack-Resistance Sensing Material. Our contact vibration sensor is designed based on a novel crack-resistance sensing material proposed in [17]. By mimicking the sensory slit organs at the leg joints of spiders, this sensing material is built by depositing a stiff, thin layer of *Pt* (less than 100nm) on a polyurethane acrylate (PUA) substrate as illustrated in Fig. 2. Transversal cracks on the *Pt* layer are formed and controlled by bending the sample during depositing. As shown in Fig. 2(a) and Fig. 3(a), when no strain is applied to the material, matching crack edges are mostly in contact with each other, resulting in a low resistance. And when extension force is applied to the sensor as shown in Fig. 2(b) and Fig. 3(b), gaps between matching crack edges increase, resulting in a higher resistance. Experiment results in [17] show that the crack-resistance sensing material is highly sensitive to vibrations and can potentially be used as a contact microphone or vibration sensor. In this work, we repeat the steps proposed in [17] to build the sensor for vibration detection. In general, the crack-resistance sensing material is equivalent to a variable resistor responsive to vibrations which we can use to simulate a common electret microphone as explained in the next section.

4.1.2 Reuse Existing Audio Hardware. We build our contact vibration sensor by tailoring the above sensing material to simulate a common electret microphone. Fig. 4(a)-(b) show the appearance and the simplified schematic of a typical electret microphone (inside the dashed box). The electret unit in Fig. 4(b) is a prepolarized dielectric material with a permanent static electric charge between its two plates. The distance between the two plates changes as a result of air pressure variation caused by sounds, which creates a voltage difference. Such voltage difference is amplified by a JFET transistor and result in a measurable voltage signal in the microphone's two terminals (*Term. 1* and *0*). For a electret microphone, the output voltage signal on *Term. 1* is always positive and below V_s. The electret microphone in the dashed box is thus equivalent to a variable resistor R_x responsive to external vibrations as shown in Fig. 4(c). We tailor the crack-resistance sensing material so that its variation of resistance in response to vibrations results in a voltage difference between *Term. 1* and *0* that matches the standard of a common electret microphone. In this way, we build a contact vibration sensor that appears exactly like a common electret microphone to the ADC unit of a COTS audio chip. By simply replacing the electret microphone with our contact vibration sensor, we can implement a vibration sensing system by reusing the existing sampling, encoding, transmission and processing in COTS audio devices.

After careful measurement and test, we match the contact vibration sensor with an electret microphone by controlling its resistance to be around 2.2kΩ when no external force is applied. This is done by tailoring the crack-resistance sensing material to the size of approximately 3mm*25mm. Fig. 5 shows the sensor built for our wearable sensing system. The gray foil is our contact vibration sensor, with conductors on both ends for signal output, and the white tape is designed to attach the sensor to the foot. The slide glass is not a part of the system.

A piezoelectric based microphone, however, cannot simply be used to replace an electret microphone because it may create a negative output voltage. Moreover, the output voltage of a typical piezoelectric sensor[1] is far beyond the output limitation of an electret microphone which is always below V_s (typically with $V_s \leq 5V$).

[1]For example, the LDT0-028K piezoelectric PVDF.

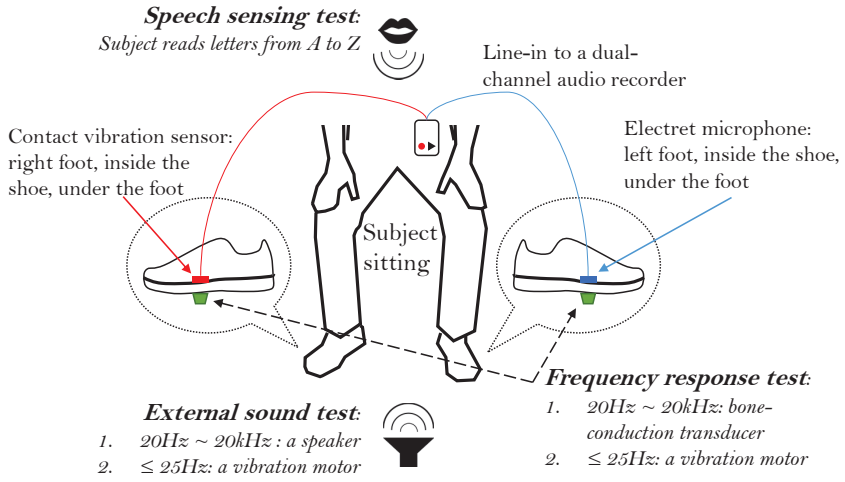


Fig. 6. Performance benchmarking setup.

As a result, fully customized systems [26] are required, making them high-cost and prohibitive for large-scale deployment. On the contrary, our sensor can be integrated into existing COTS audio devices as shown later in Sec. 5. Though *Pt* is used, the cost for each sensor is less than $1 for it is small in size and only requires a very thin layer of *Pt*.

4.2 Performance Benchmarking

In this section, we benchmark the performance of the contact vibration sensor proposed above. Fig. 6 shows the setup for performance benchmarking. Tests are conducted by asking a subject to sit still on a chair in a quiet room. A contact vibration sensor and an electret microphone are wire-connected to a dual-channel audio recorder, worn under the subject's right and left foot, respectively, inside the shoes as shown in Fig. 6. Performance benchmarking is conducted by comparing the performance of the contact vibration sensor against the electret microphone with respect to vibration *frequency response*, sensitivity to *external sounds*, and preserving privacy sensitive *speech* information.

4.2.1 Frequency Response. The frequency response test is conducted under two settings: 1) a bone-conduction transducer is attached under each shoe as shown in Fig. 6. The two transducers perform frequency sweeping from 20Hz to 20kHz simultaneously. 2) for frequency below 20Hz, we use a vibration motor with a variable rotation rate between 50rpm to 1500rpm to generate a vibration from 1Hz to 25Hz. We test the response for frequencies below 25Hz because extra low-frequency signals may contain important information such as pressure change, and can propagate through solid materials for a longer distance than high-frequency signals. Vibration data are recorded by the dual-channel audio recorder as introduced above.

Fig. 7 illustrates the comparison of frequency response of the electret microphone and our contact vibration sensor. This result suggests that both our contact vibration sensor and the electret microphone can capture vibration signals. By observing the results in Fig. 7, it is clear that our contact vibration sensor has a flatter frequency response. The electret microphone is generally more sensitive when vibration frequency is around 3kHz. However, when vibration frequency is very low (below 25Hz), our contact vibration sensor achieves better results. In this work, we focus on capturing vibration signals caused by various transportation activities under different circumstances with no predefined optimal frequency. Low-frequency signals may also be important for

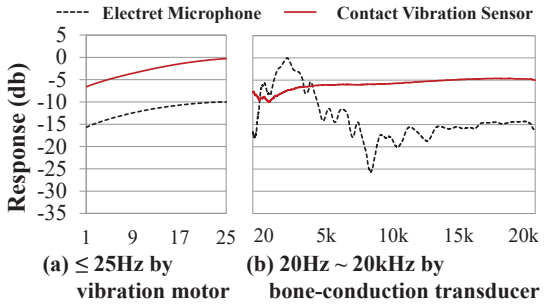


Fig. 7. Vibration frequency response.

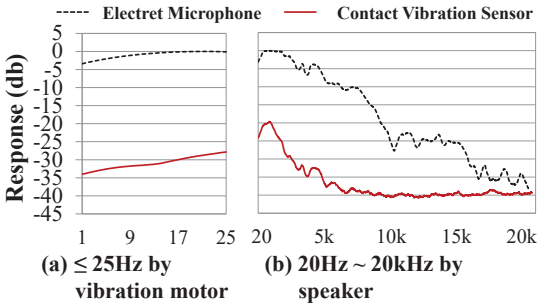


Fig. 8. Sensitivity to external sounds.

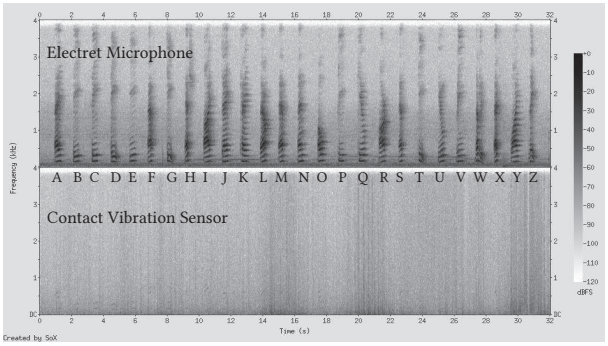


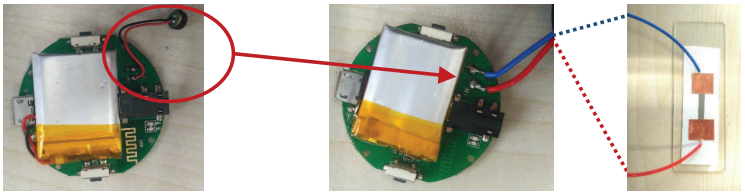
Fig. 9. Spectrogram of captured speech.

reflecting the plantar pressure changes to discriminate different activities [33]. As a result, a sensor with flatter response and being sensitive to low-frequency signals is preferred.

4.2.2 Sensitivity to External Sounds. Sensitivity to external sounds is considered harmful in our scenario because it causes the data assignment and privacy preservation issues as discussed above. To evaluate our contact vibration sensor's sensitivity to external sounds, we repeat the frequency response test and change the audio source from bone-conduction transducers to a loudspeaker as shown in Fig. 6.

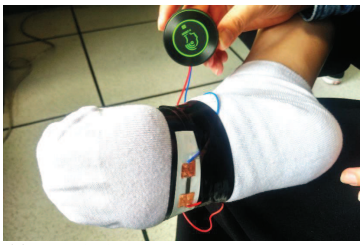
Fig. 8 illustrates the comparison of the electret microphone's and our contact vibration sensor's frequency response when exposed to external sounds from 1Hz to 25Hz (sounds of the vibration motor), and 20Hz to 20kHz (from the speaker). It is clear that our contact vibration sensor outperforms the electret microphone for being much less sensitive to external sounds. Though both the contact vibration sensor and the electret microphone are worn under the foot inside the shoe which blocks part of external sounds, a possible explanation to the above result is that the electret microphone is encapsulated in a metal shell which leaves enough space for its plates to move when air pressure changes as shown in Fig. 4(a). Our contact vibration sensor, however, is a thin foil attached to the foot which will not vibrate unless the external sounds causes the flesh or sole to vibrate.

4.2.3 Speech Sensing. Speech is often considered as sensitive information when designing audio based sensing solutions. In this test, we ask the subject to repeat letters from A to Z and record the captured vibration and audio data from both the contact vibration sensor and the electret microphone. System setup is similar to the above two tests as illustrated in Fig. 6.

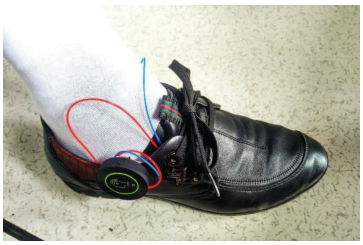


(a) The original Bluetooth audio adapter with an electret microphone

(b) Replace the electret microphone with our contact vibration sensor to build the wireless sensing system



(c) Attach the sensor to a sock to make it wearable



(d) Subject wearing the sensor under the foot inside the shoe

Fig. 10. Implementation of our Bluetooth foot-worn vibration sensing system.

Fig. 9 shows the spectrogram of captured speech data by the electret microphone (left channel) and the contact vibration sensor (right channel). As shown in the figure, the electret microphone clearly captures every letter spoken by the subject while the contact vibration sensor barely captures any signal. Though speech sounds can propagate through both the air and the flesh, the contact vibration sensor cannot capture user's speech because it is located under the foot far away from the throat and is insensitive to external sounds. As a result, we conclude that our contact vibration sensor outperforms electret microphones in privacy preserving by means of preserving users' sensitive speech information.

4.2.4 Summary. In summary, the proposed contact vibration sensor is effective for vibration sensing and outperforms electret microphones for having a flatter frequency response, resistance to external sounds, and privacy preserving. A simple experience test also reveals that our flexible sensor is more comfortable to wear than the rigid electret microphones and brass piezoelectric sensors. As a result, we conclude that the proposed sensor is suitable for our goal. We omit the performance comparison against piezoelectric sensor based solutions in this study for device simplicity and user comfort reasons, which we leave for our future work. Moreover, the shape and supporting material may also influence the sensing performance. As a pilot study, we leave the detailed discussions on these topics for our future work.

5 SYSTEM IMPLEMENTATION

We present the implementation of our wearable sensing system and the data processing pipeline.

5.1 Wearable System Implementation

Fig. 10 shows the implementation of our foot-worn vibration sensing system. We modify a COTS Bluetooth audio adapter by replacing its built-in electret microphone with our contact vibration sensor. The contact vibration sensor is attached to a sock to form a simple wearable system. The Bluetooth audio adapter has a maximum sampling rate of 8kHz which is capable of capturing vibration signals $\leq$ 4kHz.

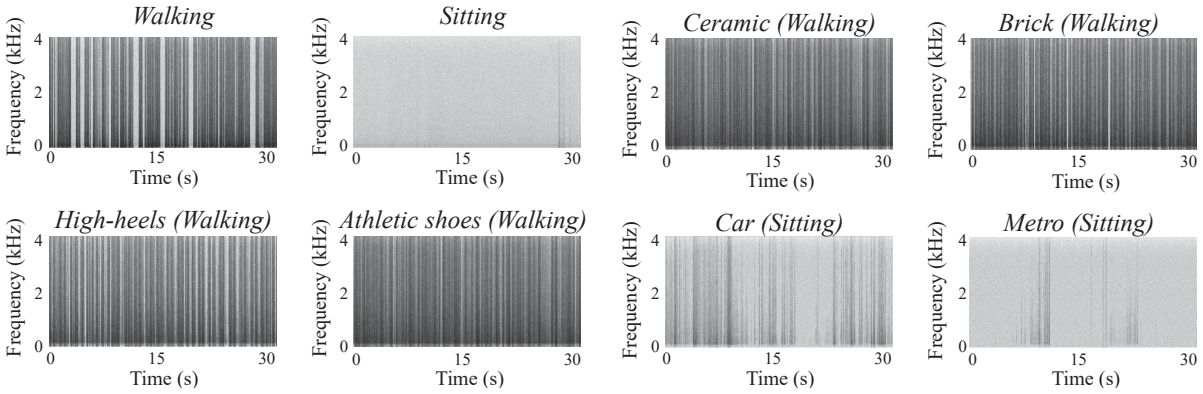


Fig. 11. Spectrograms for different transportation activities and circumstances, darker colors indicate higher amplitudes as shown in Fig. 9.

Vibration data are transmitted through Bluetooth connection to an unmodified Android smartphone. The data recording program is implemented using the audio recording programming interface provided by Android SDK. We set the audio recording format to be 8kHz 16bit Mono PCM to cope with the Bluetooth audio adapter. The received vibration data can be stored in a WAV file for offline analysis or processed online in memory according to application requirements.

5.2 Observing the Data

Before introducing the data processing pipeline, we first observe the data collected under different circumstances. Fig. 11 shows four groups of data collected with different transportation activities, wearing different shoes, walking on different types of surfaces, and sitting in different vehicles, respectively.

From Fig. 11, it is clear that different activities (*walking* vs. *sitting*) and vehicles (*car* vs. *metro*) can be discriminated by vibration signal patterns from both the frequency and time domain. Signal strength varies periodically during walking that approximately matches with each step while hardly any vibration is detected when the subject is sitting. Similarly, when sitting in different vehicles, discriminative signal patterns can be observed that approximately match with the vibration caused by the car's engine and the metro train's wheels running on railways. By comparing *high-heels* and *athletic shoes*, it can be observed that different types of shoes have similar spectral patterns during short time periods. However, they can be discriminated if patterns are extracted over a longer period. Finally, it is difficult to discriminate different road surfaces (*ceramic* vs. *brick*) simply from the spectrograms. We provide detailed analysis of the signal components and features that are important for discriminating different classes later in Sec. 6.4.2 by computing their information gain.

Summarizing the above observations, we conclude that it is possible to discriminate different transportation activities, shoes, and vehicles from the vibration signal collected. However, it is still challenging to perform accurate detection when different road surface types are involved. As a result, we adopt the audio processing and machine learning approaches [21, 26] that have shown to be promising in context detection to build our data processing pipeline.

We model the detection problem in this paper as a supervised learning problem and build the processing pipeline as shown in Fig. 12. While the frame- and window-level preprocessing steps are similar to existing audio-based event detection systems, we propose a novel classification approach by combining the independent

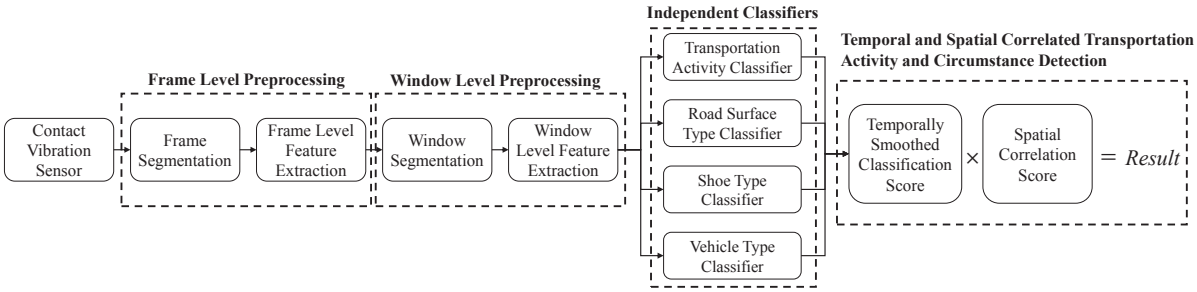


Fig. 12. Data processing pipeline.

classifiers for transportation activity and circumstance detection with respect to temporal and spatial correlations. We introduce the detailed steps involved in the data processing pipeline in the following sections.

5.3 Data Preprocessing

Because vibration is similar to audio in physical nature, we adapt the existing audio processing approaches [21, 26] for vibration-based circumstance-aware transportation activity detection.

5.3.1 Frame Segmentation. Given the input streaming data at a sampling rate of 8kHz, we apply a 50% overlapping sliding window with a fixed size to perform frame segmentation. Frame size used in existing audio-based systems varies from 23ms to 125ms [21, 23, 26]. In this work, we find the optimal frame size by testing the system's performance with frame size varies from 2ms to 128ms. After frame segmentation, we obtain a series of frames from the raw vibration data stream.

5.3.2 Frame-level Feature Extraction. We extract both time- and frequency-domain features to characterize the vibration data in each frame following existing audio processing approaches [21, 26]. Table 1 lists the frame-level features extracted. For time-domain features, we extract the signal's *energy* and *zero crossing rate (ZCR)* to characterize the signal's time-domain energy level and diversity. For frequency-domain features, we first perform FFT to obtain the signal's frequency-domain representation. We then compute features including *spectral centroid*, *spectral variance*, *spectral mean*, *spectral flux*, *spectral entropy*, *relative spectral entropy*, *spectral skewness*, *bandwidth*, *spectral rolloff* 25%-90%, *spectral slope*, and *spectral kurtosis*. We also compute the signal's energy in eight *sub-bands* with frequency ranges $(0, f_s/256)$, $(f_s/256, f_s/128)$, $(f_s/128, f_s/64)$, $(f_s/64, f_s/32)$, $(f_s/32, f_s/16)$, $(f_s/16, f_s/8)$, $(f_s/8, f_s/4)$, and $(f_s/4, f_s/2)$, and the 12 dimensional *Mel Frequency Cepstral Coefficients (MFCC)* with 20 filters. For each frame, a 36 dimensional feature vector is extracted from the raw vibration data.

5.3.3 Window-level Processing. After frame-level segmentation and feature extraction, the raw input vibration stream is converted into a sequence of frame feature vectors. We then apply a much longer sliding window with 50% overlapping to segment the frame sequence into windows. Existing work on transportation mode detection varies largely in their window sizes. Some work uses short windows of 5s [6] while others adopt large windows like 60s [11] or 120s [38] to even longer and unbounded segments [41]. Since there does not exist a commonly agreed optimal window size, we test the system's performance with window sizes from 5s to 120s increased by 5s in each step to search for the optimal window size. Because of the time cost of experimenting on the complete data set is prohibitive, we carefully choose a smaller subset to benchmark the system's performance and select the optimal set of parameters in Sec. 6.2.

For each window, we apply 10 statistical functions, i.e., *mean*, *variance*, *max*, *min*, *median*, *1st* and *3rd quartiles*, *skewness*, *kurtosis*, and *slope* to each frame-level feature to extract window-level features. After frame- and

Table 1. Frame-level features extracted (with abbreviations when applicable).

Group	Feature	Description
Time-domain	Energy	RMS of signal amplitude within a frame [21, 26]
	Zero Crossing Rate (**ZCR**)	Time-domain signal diversity of vibration data within a frame [21, 26]
Frequency-domain	Spectral Centroid (**SC**)	Center of mass across frequencies [21, 26]
	Spectral Variance (**SV**)	Energy variance across different frequencies [26]
	Spectral Mean (**SM**)	Mean energy across different frequencies
	Spectral Flux (**SF**)	Degree of signal change between frames [26]
	Spectral Entropy (**SE**)	Spectral entropy computed by FFT amplitudes
	Relative Spectral Entropy (**RSE**)	Difference between the current frame and previous frames [21]
	Spectral Skewness (**SS**)	Skewness of spectral distribution [26]
	Bandwidth	Width of the range of the frequencies that the signal occupies [21]
	Spectral Rolloff 90% (**SRF90**)	Frequency bin below which 90%, 75%, 50%, and 25% of the distribution is concentrated [26]
	Spectral Rolloff 75% (**SRF75**)	
	Spectral Rolloff 50% (**SRF50**)	
	Spectral Rolloff 20% (**SRF20**)	
	Spectral Slope (**SP**)	The shape of spectra [26]
	Spectral Kurtosis (**SK**)	
	Sub-band Energy (**SubEng[1-8]**)	Energy of 8 different frequency sub-bands [26]
	MFCC (**MFCC[1-12]**)	12-dimensional Mel Frequency Cepstral Coefficients [26]

window-level segmentation and feature extraction, we obtain a 360-dimensional feature vector to characterize vibration data within a window.

5.3.4 Frame Admission Control. By observing the collected data in Fig. 11, it is clear that in many cases (e.g., *sitting*, *car*, and *metro*), a large portion of data contain no significant vibration signals, i.e., silent periods. Frames obtained from these silent periods often contain no interesting information and are similar to each other. Processing these silent frames may cost valuable computational and battery power, which reduces the system's efficiency. As a result, in this paper, we propose a simple frame admission control approach by filtering out the low energy frames obtained from silent periods. We empirically determine a reference low energy level $energy_{ref}$ by the cutting energy level below which 1% of the frames in the training data set reside. Then the energy threshold $energy_{th}$ to determine a frame to contain no interesting information is computed as:

$$energy_{th} = \alpha \cdot energy_{ref}$$

where $\alpha \geq 0$ is a scaling factor.

Frames with energy below $energy_{th}$ are filtered to skip the above feature extraction process. However, to cope with the window-level processing step, these frames cannot be simply discarded. Instead, a predefined frame feature vector that represents all low energy frames is used to replace the filtered frames.

5.4 Temporal and Spatial Correlated Classification

In this work, we model the transportation activity and circumstance detection problem as a supervised multi-label classification problem. Given the continuous stream of window-level features, the classification problem is to assign a label to each type of the activity and circumstances. A naïve approach to solve this problem is to build four independent classifiers for *activity*, *road surface*, *shoe*, and *vehicle* types. However, this naïve approach fails to explore the intrinsic nature of the data: 1) the activities and circumstances often remain stable in temporally adjacent windows, e.g., people walking down a street is unlikely to change the activity or shoes frequently; 2) different combinations of activities and circumstance types in the same window have different possibilities, e.g., it is unlikely that one will jog on the plastic tracks wearing high-heels. As a result, in this work, we propose to

perform transportation activity and circumstance detection with temporal and spatial correlated classification. We note here that the main contribution of this work is the design of the novel contact vibration sensing system. The following classification framework is mainly proposed to better demonstrate the effectiveness of our system on the task. Further topics such as the optimal choices of classifiers are out of the scope of this paper.

5.4.1 Label Distribution from Independent Classifiers. First, we select a base classifier and build four instances for transportation activity, road surface, shoe, and vehicle type detection, respectively. In this work, we propose to use the Random Forest Classifiers [5] which have shown to be promising in Wi-Fi signal strength- [25] and capacitor-based [22] transportation mode detection systems to build our models. We evaluate the performance of the proposed Random Forest-based model in Sec. 6. Moreover, in Sec. 6.6.2, we compare the proposed model with two extensions of HMMs [4, 10, 37] which are also frequently used for complex activity recognition.

Instead of directly using the classification results, we obtain the label distribution from each individual classifier independently given the current window.

Definition 5.1. Label distribution. Given the current window at time t, the label distribution of the i-th classifier ($i = 1, 2, 3, 4$ corresponding to transportation activity, road surface, shoe, and vehicle type, respectively) is a vector

$$\mathbf{D}_{i,t} =< d^j_{i,t} >$$

where $d^j_{i,t}$ is the classifier's confidence on classifying the current window into label j ($j = 1, ..., N_i$ where N_i is the number of possible labels for the i-th classifier).

We use Weka [7] to implement the Random Forest Classifier. As a result, the label distribution can be obtained directly from the classifier which is computed by normalizing the distribution of each random tree.

5.4.2 Temporally Smoothed Classification Score. Based on the label distribution from each individual classifier, we explore the temporal correlation among adjacent windows by computing the temporally smoothed label distribution as follows.

$$\tilde{d}^j_{i,t} = \frac{1}{Z_{i,t}} \sum_{n=0}^{N-1} \frac{1}{n+1} d^j_{i,t-n}$$

where $\frac{1}{n+1}$ is the weight of window $t-n$ when computing the temporally smoothed label distribution for window t, N is the number of windows involved (we empirically set $N = 10$ based on a preliminary testing result), $Z_{i,t} = \sum_{j=1}^{N_i} \sum_{n=0}^{N-1} \frac{1}{n+1} d^j_{i,t}$ is the normalizing factor that keeps $\sum_{j=1}^{N_i} \tilde{d}^j_{i,t} = 1$.

Let $\mathbf{L} =< l_i >, i = 1, ..., 4$ be a combination of labels for transportation activity, road surface, shoe, and vehicle types, where l_i is the label assigned to the i-th class. We define the temporally smoothed classification score of window t for label combination $\mathbf{L}$, $C_t(\mathbf{L})$, as the sum of temporally smoothed label distribution of $l_i \in \mathbf{L}$ as

$$C_t(\mathbf{L}) = \sum_{i=1}^{4} \tilde{d}^{l_i}_{i,t} \tag{1}$$

5.4.3 Spatial Correlation Score. When considering the spatial correlations among different transportation activities and circumstances, we aim at eliminating combinations that are unlikely to happen, e.g., jogging wearing high-heels. The basic idea behind the proposed spatial correlation score $\mathcal{S}(\mathbf{L})$ is to approximate the joint density of $\mathbf{L}$. A simple way to obtain the spatial correlation score is to count the support of different label combinations in the training data set. However, the simple counting approach is biased to combinations that appear frequently in the training data set. To solve this problem, we propose the following piecewise function for the

Table 2. Information of the subjects.

Subject No.	Gender	Age	Height (cm)	Weight (kg)
1	Male	21	170	73
2	Female	22	168	52
3	Male	22	172	62
4	Male	33	177	70
5	Female	60	163	60
6	Male	60	173	68

spatial correlation score:

$$\mathcal{S}(\mathbf{L}) = \begin{cases} supp(\mathbf{L}), & if \quad 0 \leq supp(\mathbf{L}) < threshold \\ 1, & if \quad supp(\mathbf{L}) \geq threshold \end{cases} \tag{2}$$

where $supp(\mathbf{L})$ is the support of label combination $\mathbf{L}$ in the training data set, and $threshold$ is a small real number (empirically set to 1‰) to eliminate the unlikely combinations while keeping all the possible combinations.

5.4.4 Temporal and Spatial Correlated Classification. Based on the above classification and spatial correlation scores, we define the final synthesized score of label combination $\mathbf{L}$ for window t, $\mathcal{F}_t(\mathbf{L})$, as follows:

$$\mathcal{F}_t(\mathbf{L}) = C_t(\mathbf{L}) \cdot \mathcal{S}(\mathbf{L}) \tag{3}$$

where $C_t(\mathbf{L})$ and $\mathcal{S}(\mathbf{L})$ are as defined in Equ. (1) and (2), respectively.

Given a series of window-level features and the current window at time t, the temporal and spatial correlated classification approach for transportation activity and circumstance detection is done by finding the label combination with the highest final score as follows:

$$result_t = \arg\max_{\mathbf{L}} \mathcal{F}_t(\mathbf{L}) \tag{4}$$

As shown later in Sec. 6.6, the proposed temporal and spatial correlated classification approach is effective in recovering the detection errors made by the independent classifiers.

6 EVALUATION

We evaluate the performance of *SpiderWalk* system in this section.

6.1 Data Collection and Methodology

6.1.1 Subjects and Data Collection. With IRB approval, data collection is done with six subjects—two females and four males—over a month. Table 2 lists the subjects' information. The subjects are asked to wear the foot-worn sensor as shown in Fig. 10 during their daily livings. Vibration data are transmitted wirelessly through Bluetooth connections to smartphones on which we deploy the data collection program. Data transmission and formatting are implemented using the existing audio recording framework provided by the Android OS as introduced in Sec. 5. The ground truth for the activity, road surface, vehicle, and shoe types are manually labeled by the subjects.

It is important to note that our system can only function when the sensor touches the ground, especially when trying to capture the vibration patterns to discriminate different vehicles. As a result, we ask the subjects to keep their feet wearing the sensors fully touching the ground when sitting or standing.

6.1.2 The Collected Data Set. Table 3 summarizes the types of activities and circumstances collected in our data set and the distribution of data hours across subjects and labels. By the time this paper is written, we have collected a total amount of approximately 220 hours of data. The *null* class in the road surface category

Table 3. Labels of activities and circumstances collected & data distributions across subjects and labels.

Subject No.	Hours	Activity	Hours	Road Surface	Hours	Shoe	Hours	Vehicle	Hours
1	18.2	Walking	120	Ceramic	20.7	Dress Shoes	86.3	Bus	38.2
2	13.2	Running	6.2	Grass	10.9	Athletic Shoes	93.4	Metro	13.9
3	16.8	Riding	21.5	Wooden	32.5	Slippers	22.6	Bike	21.5
4	35	Idle	71.9	Tar	32.4	High Heels	17.2	Car	7
5	60.6			Brick	10.6			On Foot	138.9
6	75.8			Plastic	6.1				
				Mud	25.7				
				Null	80.5				

represents cases the subject is not traveling on foot. And the *on foot* class in the vehicle category represents cases the subject is not traveling by vehicles. While the total number of possible label combinations of the above four categories is 640, there are 61 label combinations in the collected data set. The collected label combinations are much fewer for two reasons: 1) **mutually exclusive**: some labels that are mutually exclusive by definition, e.g., *road surface (null) + vehicle (on foot)* is not a valid combination as explained above; 2) **unlikely combinations**: some combinations are unlikely to happen in real-life, e.g., *activity (running) + vehicle (car)*, and *activity (running) + shoe (high heels)*. We do not ask the subjects to perform unnatural or risky activities even if some combinations such as *activity (running) + shoe (high heels)* are possible.

Another observation made from Table 3 is that the data distribution across different subjects and labels is imbalanced. This is partially because we ask the subjects to collect data in a close to nature manner. As a result, we obtain fewer data for *running* than *walking*. Additionally, subject 1 to 3 spend most of their time in classrooms and labs which results in a large amount of *idle* data. We removed most of these *idle* data to balance the data distribution because the vibration signals in this class are very sparse and hardly contain any useful information. While the data imbalance issue may potentially affect the system's performance, we discuss this issue in Sec. 6.3.3 in detail. Though the data is imbalanced, the collected data set is feasible for performance evaluation as suggested by the results in Sec. 6.3.3.

6.1.3 Experiment Design and Organization. We start our experiments by selecting the optimal set of parameters including frame and window sizes, and the number of trees in the Random Forests through benchmarking in Sec. 6.2. We use the detection accuracy for different categories and the average detection accuracy to evaluate the system's performance.

We then use the selected parameters to evaluate the system's performance in Sec. 6.3, in which we also discuss issues related to the collected data set including the impact of physical factors and the data imbalance issue. Detailed performance is presented through confusion matrices, precision, recall, and F1 score.

In Sec. 6.4, we evaluate the performance of system components including the frame admission control and feature selection. Together with the results presented in Sec. 6.5.1, we gain an in-depth understanding of the impact of these components on detection accuracy and system overheads. By computing the information gain for different features during the feature selection process, we gain some insights into the correlations between the labels and their preferred features.

Sec. 6.5 evaluates the system's overheads on the smartphone and the sensor node with respect to CPU, memory, and battery power consumptions. We compare the performance of the proposed system with a naïve, independent classifier model, extensions of HMMs, and by using different sensing modalities in Sec. 6.6. Finally, Sec. 6.7 provides some discussions on the system's design and implementation issues.

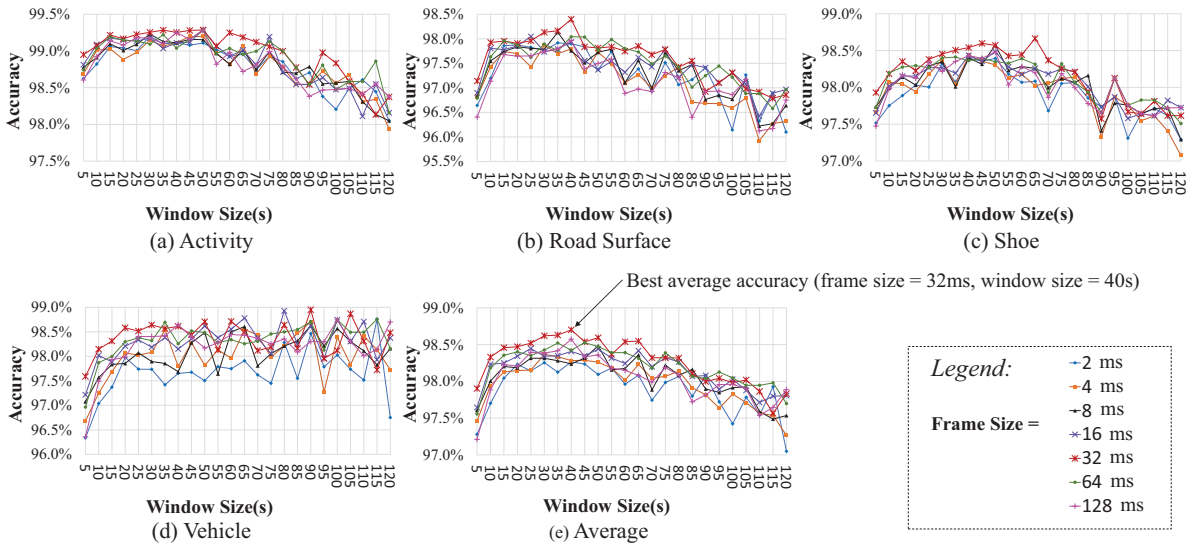


Fig. 13. Detection accuracy with different frame and window sizes (better viewed in color).

6.2 System Parameter Benchmarking

In this section, we test the system's performance with different parameters to discover the optimal parameter set. During these tests, system components such as admission control and feature selection are not enabled. Additionally, we test a large number of combinations for frame and window sizes, as well as the number of trees in this experiment. The time cost of testing on the complete data set of 220 hours is prohibitive. As a result, we obtain the optimal system parameters by testing the system's performance on a benchmark data set containing 10% of the total data. The benchmark data set is selected by matching the trend of performance change with the results obtained using the complete data set on a few samples in the parameter space.

6.2.1 Frame and Window Sizes. We first evaluate the system's detection accuracy with different frame sizes (i.e., 2ms to 128ms) and window sizes (i.e., 5s to 120s with a step of 5s). During the test, we set the number of trees in the Random Forest models to 20. Fig. 13(a)-(d) show the detection accuracy for activity, road surface, shoe, and vehicle types, respectively. Fig. 13(e) summarizes the results by averaging the above detection accuracies.

A general observation from the figures is that our system's performance is not sensitive to frame and window sizes. And the optimal frame and window sizes for different categories are close to each other. The optimal frame size of 32ms is agreed by all four categories. The optimal window sizes for activity, road surface, shoe and vehicles are 50s, 40s, 65s, and 90s, respectively. From the results of average detection accuracy, the optimal frame and window sizes are decided to be 32ms and 40s, respectively. In Sec. 6.3, we report the system's detailed performance on the complete data set under this setting. We use this set of parameters in the following experiments unless noted otherwise.

6.2.2 Number of Trees. One advantage of using Random Forests as the classification models is their flexibility of choosing different number of trees to achieve a trade-off between performance and cost. In this test, we evaluate the system's performance with different number of trees on the benchmark data set. During this test, we set the frame size to be 32ms and window size to be 40s following the optimal values obtained from the above experiment. Fig. 14 illustrates the detection accuracy with different number of trees from 1 to 150. There is a sharp

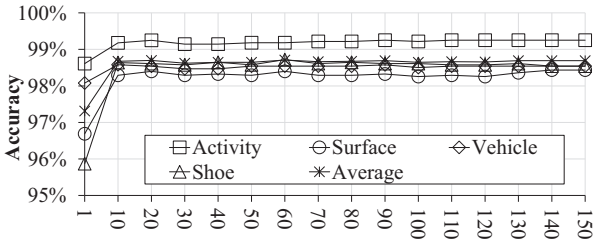


Fig. 14. Detection accuracy with different number of trees.

increase in detection accuracy when the number of trees increases from 1 to 20. After the detection accuracy reaches 98.7% with 20 trees, the increase in tree number does not show a clear positive effect on improving the system's detection accuracy. On observing that the time cost for training and testing the models and the memory cost to store the models increase linearly with the number of trees, we conclude that the best trade-off between accuracy and resource usage is to set the number of trees to 20.

6.3 Performance Evaluation

Through the above preliminary experiments, we select the optimal system parameters to be frame size of 32ms, window size of 40s, and the number of trees to be 20. In this section, we evaluate the system's performance on the complete data set using these parameters. We first present the detailed detection performance on the complete data set. We then study the impact of different physical factors including gender, age, height, and weight on the system's performance. Finally, we study the data imbalance issue in the last experiment.

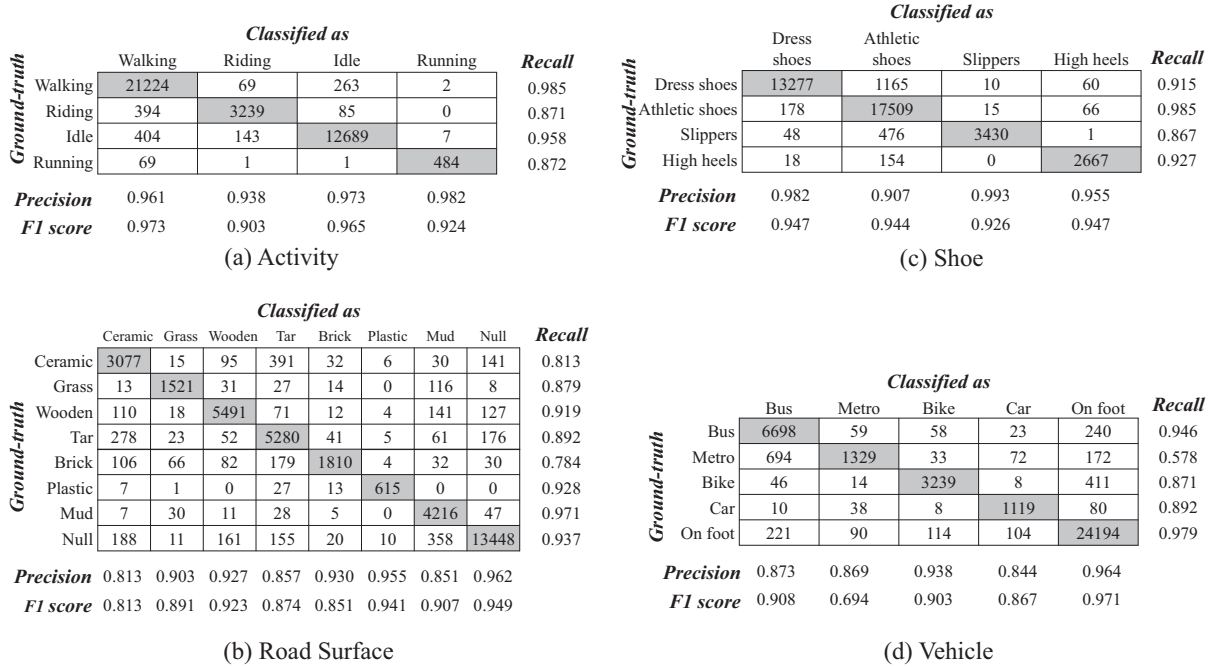
6.3.1 Detailed Detection Performance. We evaluate the system's performance using the optimal parameter set on the complete data set by 10-fold-cross-validation. And the overall result suggests an average detection accuracy of 93.8%. We break down the results into details by presenting the confusion matrices and computing the precision, recall, and F1 score for each category as illustrated in Fig. 15.

From Fig. 15(a), the result shows that the average precision, recall, and F1 score for transportation activity detection are all 0.96, suggesting our system can achieve accurate detection of various activities. Both of the *riding* and *running* activities have the lowest recall of 0.87. Detailed result shows that 10.6% and 12.6% of the *riding* and *running* instances are recognized as *walking*, respectively. A possible explanation is that in some cases the user is *riding* or *running* (*jogging*) slowly which results in a vibration pattern similar to *walking*.

Next, from Fig. 15(b), the average precision, recall, and F1 score for road surface detection are all 0.91. The lowest precision of 0.81 and recall of 0.78 are from surface types *ceramic* and *brick*, respectively. This is because 10.5% instances of *ceramic* are recognized as *tar*, and 12.3% instances of *brick* are recognized as *ceramic* or *tar*. Possible reasons are ceramic tiles are rough and similar to tar roads, and we mark roads covered with marble tiles which are very similar to ceramic tiles as brick roads, making it difficult to distinguish from each other.

From Fig. 15(c), the average precision, recall, and F1 score for shoe type detection are 0.95, 0.94, and 0.94, respectively. *Slippers* has the lowest recall of 0.87 because 12% of the instances are mistakenly classified as *athletic shoes*. It can possibly be explained by the data imbalance issue because the *athletic shoes* class has much more instances than the *slippers* class, which we discuss in detail in Sec. 6.3.3.

Third, Fig. 15(d) suggests our system can detect different vehicle types with average precision, recall, and F1 score of 0.94, 0.94, and 0.93, respectively. The lowest precision of 0.84 and recall of 0.58 are from vehicle types *car* and *metro*, respectively. Further study shows that 8% of the instances classified as *car* are actually *on foot*. And 30% of the instances of *metro* are classified as *bus*. A possible explanation for the low precision of *car* is



Classified as

Ground-truth	Walking	Riding	Idle	Running	*Recall*
Walking	21224	69	263	2	0.985
Riding	394	3239	85	0	0.871
Idle	404	143	12689	7	0.958
Running	69	1	1	484	0.872
Precision	0.961	0.938	0.973	0.982	
F1 score	0.973	0.903	0.965	0.924	

(a) Activity

Classified as

Ground-truth	Dress shoes	Athletic shoes	Slippers	High heels	*Recall*
Dress shoes	13277	1165	10	60	0.915
Athletic shoes	178	17509	15	66	0.985
Slippers	48	476	3430	1	0.867
High heels	18	154	0	2667	0.927
Precision	0.982	0.907	0.993	0.955	
F1 score	0.947	0.944	0.926	0.947	

(c) Shoe

Classified as

Ground-truth	Ceramic	Grass	Wooden	Tar	Brick	Plastic	Mud	Null	*Recall*
Ceramic	3077	15	95	391	32	6	30	141	0.813
Grass	13	1521	31	27	14	0	116	8	0.879
Wooden	110	18	5491	71	12	4	141	127	0.919
Tar	278	23	52	5280	41	5	61	176	0.892
Brick	106	66	82	179	1810	4	32	30	0.784
Plastic	7	1	0	27	13	615	0	0	0.928
Mud	7	30	11	28	5	0	4216	47	0.971
Null	188	11	161	155	20	10	358	13448	0.937
Precision	0.813	0.903	0.927	0.857	0.930	0.955	0.851	0.962	
F1 score	0.813	0.891	0.923	0.874	0.851	0.941	0.907	0.949	

(b) Road Surface

Classified as

Ground-truth	Bus	Metro	Bike	Car	On foot	*Recall*
Bus	6698	59	58	23	240	0.946
Metro	694	1329	33	72	172	0.578
Bike	46	14	3239	8	411	0.871
Car	10	38	8	1119	80	0.892
On foot	221	90	114	104	24194	0.979
Precision	0.873	0.869	0.938	0.844	0.964	
F1 score	0.908	0.694	0.903	0.867	0.971	

(d) Vehicle

Fig. 15. Detection result breakdown.

some instances of sitting in the car is collected when the car is not moving or moving slowly, making it easy to be confused with the *idle* class whose corresponding vehicle type is *on foot*. And the low recall rate of *metro* is possibly due to the data imbalance issue because *bus* has two times more instances than *metro*, and the vibration patterns of these two vehicles are sometimes similar. Another observation made from the results is that the system cannot effectively detect whether the user is walking in a motorized vehicle (*bus* or *metro*), or on the road. It mis-classifies all the instances of *walking in the metro* to *walking on the roads*, and 90% of the instances of *walking in the bus* to *walking on the roads*. A possible explanation is the subtle vibrations from the vehicles are overwhelmed by the vibration and pressure signals caused by walking.

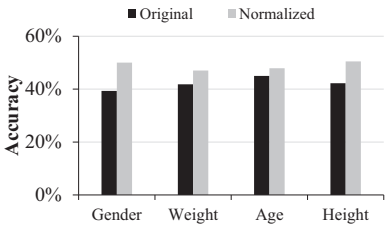
In summary, our system achieves an average detection accuracy of 93.8%, suggesting the proposed sensing and detection system is effective in identifying different transportation activities and their circumstances.

6.3.2 Impact of Physical Factors. As shown in Table 2, the six subjects vary in physical factors including gender, age, height, and weight. In this section, we conduct experiments to understand how these physical factors impact the performance of our system. For each of the four factors listed above, we group the subjects into two groups as shown in Fig. 16(a). We perform cross-group evaluation by using one group for training and the other group for testing. The average detection accuracy is used to evaluate the system's performance.

Fig. 16(b) illustrates experiment results. The black bars show the average cross-group detection accuracy with respect to different physical factors drop to around 40%. By comparing to the results presented in Sec. 6.3 above, we conclude that the system is dependent on physical factors. A closer study reveals that the physical factors affect the detection of different categories differently. The activity and vehicle types are less affected for remaining a detection accuracy of above 60% and 50%, respectively. The shoe and road surface types suffer more from factor differences with accuracies drop to around 30%. Possible explanations include: 1) the sensors are not

	Gender *Male vs. Female*	**Weight** *> 65kg vs. < 65kg*	**Age** *< 25 vs. ≥ 60*	**Height** *≤ 170 cm vs. > 170 cm*
Group 1 (Subject No.)	1,3,4,6	1,4,6	1,2,3	1,2,5
Group 2 (Subject No.)	2,5	2,3,5	5,6	3,4,6

(a) Group subjects according to physical factors



(b) Experiment results

Fig. 16. Impact of physical factors on the system's performance.

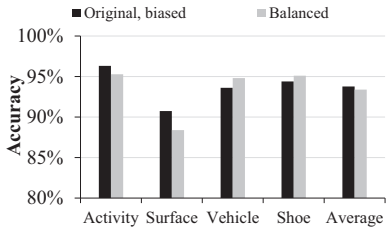


Fig. 17. Performance on the balanced data set.

calibrated to have the same response level to vibrations; 2) the subjects are wearing different shoes even if they belong to the same class, e.g., dress shoes for males and females are quite different, which further affects the detection of road surfaces.

We make a simple attempt to deal with the above issues by normalizing the window-level features. And the results shown by the gray bars in Fig. 16(b) suggest a 7% improvement in accuracy on average. Although the above experiment results suggest our system is dependent to physical factors, we demonstrate the potential of improving the system's performance by a simple technique. Further directions include calibrating the sensors, collecting more training data, building more optimized models, and etc, which we leave for our future work.

6.3.3 Data Imbalance Issue. As shown in Table 3, the distribution of data is imbalanced across different subjects and labels. And the results presented above in Sec. 6.3 also suggest the detection performance of the *metro* and *slippers* is affected by the imbalance issue. However, the impact of data imbalance on the system's performance is complicated. On the one hand, the spatial correlations modeled in our classification framework may be overfitting the distribution in the training data. As a result, the detection results may bias to labels with more instances. On the other hand, if such bias really exists in the real-world, it is expected to improve the performance of the system.

To study this issue, we obtain a balanced data set by carefully resampling the original, biased data set following two criteria: 1) we balance the number of instances for each label combination to approximately 240 instances; 2) for each label combination, we select instances evenly from subjects providing the data to eliminate the imbalance among subjects. We compare the results obtained from the original, biased data set against the results from the balanced data set. The results illustrated in Fig. 17 suggest the two data sets achieve comparable results with the performance on the balanced data set a little lower on average accuracy by 0.37%. For some categories, including activity and road surface, the system's performance on the original data set is better than on the balanced data set by 1%, and 2.5%, respectively. For the other two categories, the balanced data set outperforms the original data by 1.2% and 0.8% for vehicle and shoe type detection, respectively.

Combining the results presented in this experiment and in Sec. 6.3, we conclude that the data imbalance issue has an impact on the system's performance, especially for vehicle and shoe detection. However, such impact is insignificant in general. Several conclusions can be drawn from the results. On the one hand, it suggests the results on the complete data set are valid because the data imbalance issue does not have a decisive impact on the results. On the other hand, it also suggests the spatial correlations modeled in our classification framework is not as important as expected. A possible explanation is that the individual classifiers are carefully tuned during benchmarking in Sec. 6.2. And when the parameters are less optimized, the temporal and spatial correlations can effectively improve the results by providing chances to fix the detection errors made by the individual classifiers as shown later in Sec. 6.6.1.

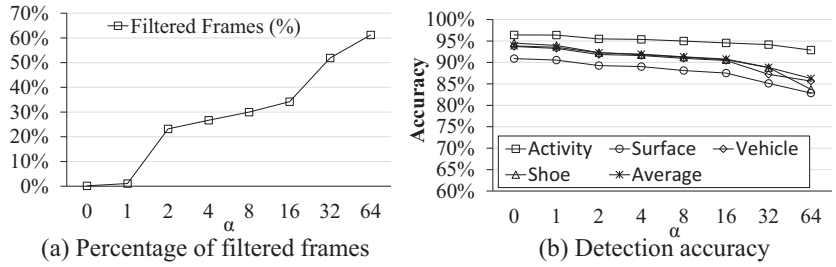


Fig. 18. Admission control performance with different α values.

Table 4. Examples of features with the highest information gain for different classes. Features are presented in the form $WindowLvFunction\text{-}FrameLvFeature$ ($InformationGain$). Abbreviations for frame level features are listed in Table 1.

Class	**Features with the highest information gain**
Idle **vs.** *Other Activities*	Var-MFCC5 (0.47), Var-MFCC6 (0.47), Var-MFCC3 (0.47), Var-MFCC4 (0.47), Var-MFCC8(0.46)
Walking **vs.** *Others Activities*	Var-SubEng7 (0.38), Var-SubEng6 (0.38), Var-SubEng8 (0.37), Max-MFCC1 (0.37), Mean-SubEng7 (0.37)
High heels **vs.** *Other Shoes*	Q1-MFCC8 (0.23), Q1-MFCC10 (0.22), Med-MFCC11 (0.21), Q3-RSE (0.21), Q1-MFCC6 (0.20)
Athletic shoes **vs.** *Other Shoes*	Min-SRF75 (0.08), Min-SRF90 (0.07), Min-MFCC2 (0.07), Min-MFCC3 (0.06), Min-MFCC4 (0.06)
Bus **vs.** *Other Vehicles*	Q1-MFCC11 (0.39), Q1-MFCC9 (0.35), Min-MFCC3 (0.34), Q1-MFCC11 (0.31), Q1-MFCC7 (0.30)
Car **vs.** *Other Vehicles*	Med-MFCC11 (0.06), Med-MFCC8 (0.06), Med-MFCC10 (0.06), Med-MFCC6 (0.06), Q1-MFCC4(0.06)
Grass **vs.** *Others Road Surfaces*	Q3-SubEng2 (0.04), Med-SubEng2 (0.04), Med-SubEng3 (0.04), Q3-SubEng3 (0.04), Skew-ZCR (0.04)
Mud **vs.** *Others Road Surfaces*	Skew-RSE (0.15), Mean-RSE (0.15), Q3-RSE (0.15), Med-RSE (0.15), Max-RSE (0.15)

6.4 System Components

We evaluate the performance of system components in this experiment.

6.4.1 Frame Admission Control. We evaluate the performance of frame-level admission control by setting different scaling factors (α). The result is illustrated in Fig. 18. The baseline is set to $\alpha = 0$ with admission control disabled. We then increase α by powers of two (i.e., 2^n) starting with $n = 0$. When $\alpha = 1$, 1% of the frames are filtered and the resulting detection accuracy is similar to the baseline with a slight increase. For $\alpha = 2$, 23% of the frames are filtered and the detection accuracy drops for 1.6% on average. By further increasing α, the percentage of frames filtered increase linearly with n and the accuracy also decreases. When $\alpha = 64$, 61.2% of the frames are filtered and the average detection accuracy drops by over 7%.

It can be observed from Fig. 18(b) that by increasing α to 64, the detection accuracy for shoes experiences more decrease (approximately 10% drop) than other categories. A possible explanation is that different shoes are discriminated by the differences in the subtle vibrations from the soles. By increasing α from 32 to 64, frames containing subtle vibrations but with low overall energies are filtered, leading to the sharp drop in detection accuracy for shoes. On the contrary, the loss in accuracy for activity detection is relatively less significant (approximately 3.5% drop) by the mean time. This result suggests activity detection possibly rely more on the variance of general signal patterns rather than the subtle vibration patterns in each frame.

In summary, we conclude that frame admission control will decrease the performance of the system, which should be used carefully. For our data set, the result suggests that the largest α is 16 when 34.2% frames are filtered while keeping the average detection accuracy to be above 90%.

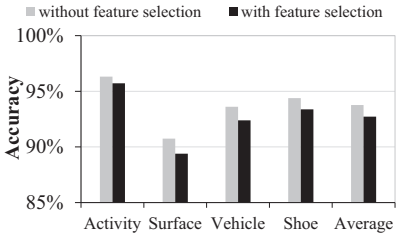


Fig. 19. Feature selection performance.

Table 5. System overheads.

Configuration	CPU	Memory	Battery
Baseline	11% - 40%	≤42MB	660mW
Admission Control 10%	11% - 38%	≤41MB	555mW
Admission Control 20%	10% - 33%	≤38MB	592mW
Admission Control 30%	10% - 33%	≤35MB	520mW
Admission Control 40%	7% - 30%	≤33MB	443mW
Admission Control 50%	7% - 30%	≤30MB	433mW
Feature Selection	7% - 30%	≤32MB	473mW

6.4.2 Feature Selection. Extracting and classifying the 360-dimensional feature vector is prohibitive for the smartphones for its CPU, memory, and battery power consumptions. As a result, we propose to use feature selection techniques to reduce the feature numbers. To select a small subset of features with strong discriminative power, we compute the information gain of each feature to measure its ability to classify the instances. We treat the problem as a series of binary classification problems when computing the information gain. For example, given the instances of different types of shoes, we compute the information gain of each feature when discriminating one type of shoes against all the other types of shoes. The same approach is applied to activities, road surfaces, and vehicles.

Table 4 lists some examples we obtained from the results. The complete list of detailed results is omitted due to page limits. The table suggests for the *idle* activity, the combinations of window level function *variance* with different MFCC features are the most discriminative features. For the *walking* activity, the preferred window level functions and frame level features are more diverse. Frame level features of high-frequency *Sub-band Energy* (e.g., SubEng6-8) are more important than other features. For the *high heels*, frame level MFCC features corresponding to the high-frequency components of the signal (e.g., MFCC8, MFCC10, and MFCC11) are more important compared to the *athletic shoes* which prefers SRF and low-frequency frame level features such as MFCC2, and MFCC3. For different vehicles, MFCC features have shown to be important for both the *bus* and *car*. However, the corresponding window level functions are largely different for the two vehicles. Finally, for different road surfaces, low-frequency *Sub-band Energy* (e.g., SubEng2-3) are important frame level features for the *grass*, while the RSE features dominates the most important frame level features for the *mud* class.

Based on the above analysis, we select a feature subset by combining the features with the top ten highest information gain for different classes. The final selected window-level feature vector contains 85 features. Compared to the original 360-dimensional feature vector, the selected feature vector is approximately 24% in size. From Fig. 19, it is clear that the selected feature set achieve similar result (1.1% drop in accuracy on average) compared to the original one with much fewer features extracted. Further studies presented next in Sec. 6.5.1 suggest the system's overheads are reduced by feature selection.

6.5 System Overheads

In this section, we evaluate the system's overheads on CPU, memory, and power consumptions on the collection and processing node (i.e., the smartphone), and the battery consumption of the sensor.

6.5.1 Smartphone's Resource Overheads. In this experiment, we evaluate the system's overheads on computing, memory, and battery power under different configurations. The test is conducted on a Google Nexus 5

Table 6. Accuracy comparison of different approaches.

Category	num trees = 20		num trees = 1	
	Independent	**Proposed**	**Independent**	**Proposed**
Activity	95.1%	96.3%	89.5%	94.3%
Road Surface	87.6%	90.8%	73.2%	81.2%
Shoe	92.4%	94.3%	81.6%	87.8%
Vehicle	91.6%	93.6%	84.1%	90.8%
Average	91.6%	93.8%	82.1%	88.5%

smartphone. We use the system monitor provided by Android Studio[2] to measure the system's CPU and memory costs, and use PowerTutor[3] to monitor battery power consumption.

Table 5 summarizes the system's overheads under different configurations. The baseline configuration disables both frame admission control and feature selection. For frame admission control, we set different rates of filtered frames to measure the system's overhead. For feature selection, we compare the system's overheads with the baseline configuration. When testing the overheads of one system component (i.e., frame admission control or feature selection), we disable the other component to make a clear comparison with the baseline.

The CPU usage during each test varies largely across time. Generally, there is a mild decrease in CPU usage by enabling frame admission control and feature selection, which can be observed from Table 5. Similar observations can also be made for memory usage. From Table 5, there is a steady decline in battery power cost by enabling frame admission control and feature selection. Generally, our system's overheads are comparable to the non-speech body sounds processing system proposed in [26] and GPS positioning systems on smartphones [20]. We plan to explore techniques such as using the smartphones' co-processors [9] or task offloading [24] to reduce the system's overheads.

In summary, frame admission control and feature selection have shown to be effective in reducing the system's overheads. Combining the performance testing results presented above, a reasonable system configuration is to set the frame admission control rate to be $\leq 40\%$ and enabling feature selection to balance the system's detection accuracy and overheads.

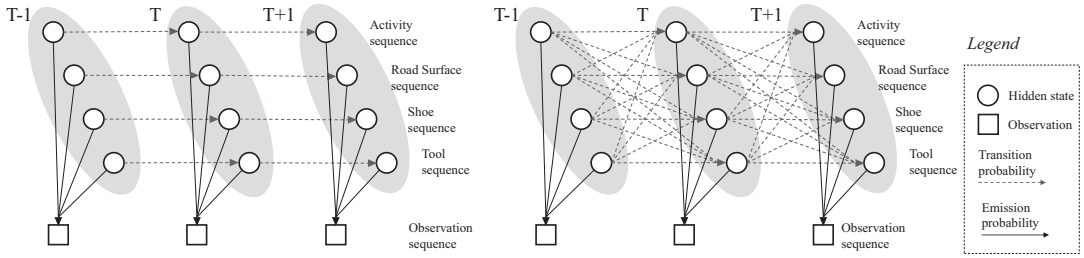
6.5.2 Sensor's Power Consumption. We evaluate the power consumption of the sensor by measuring the battery's output currency using an ammeter. The result suggests the sensor's power usage is approximately 21mA on average, and the 150mAh built-in battery can support the sensor to work continuously for 7.14 hours in theory. Real-world tests suggest the battery can actually support the sensor for 6.76 hours without recharging. The working hours of the sensor is sufficient for daily usage scenarios because we can further decrease the power consumption by turning off the sensor during the *idle* periods, which can be effectively detected by the smartphone's built-in inertial sensors as shown later in Sec. 6.6.3.

6.6 Comparison Studies

In this section, we present the results of comparison studies on three aspects: 1) compare with a naïve system design which uses four independent classifiers; 2) compare with other models including two extensions of HMMs; and 3) compare with other sensing modalities.

6.6.1 Compare with Independent Classifiers. In this experiment, we compare the performance of the proposed classification approach (temporal and spatial correlated classification) against a naïve solution that builds four independent classifiers for the four categories to show its effectiveness.

[2]https://developer.android.com/studio/index.html
[3]http://ziyang.eecs.umich.edu/projects/powertutor/



(a) Factorial Hidden Markov Models (FHMMs)

(b) Coupled Hidden Markov Models (CHMMs)

Fig. 20. DAG representations of FHMMs and CHMMs used for comparison studies.

Table 7. Accuracy comparison of different approaches.

Model	Activity	Road Surface	Vehicle	Shoe	Average
FHMMs	81.8%	58%	73.7%	72.6%	71.5%
CHMMs	88.1%	62.4%	81.1%	74.9%	76.6%
Proposed	96.3%	90.8%	94.3 %	93.6%	93.8%

Table 8. Accuracy comparison of different sensing modalities (two male subjects, without external noise).

Category	**Vibration**	**Audio**	**Inertial**	**Vibration + Inertial**
Activity	98.1%	98.8%	97.5%	97.8%
Road Surface	91.8%	94.9%	79.1%	93.1%
Shoe	98.5%	95.8%	79.3%	99.1%
Vehicle	98.4%	98.6%	91.8%	98.7%
Average	96.7%	97%	86.9%	97.2%

As shown in the left half of Table 6, when the number of trees is set to the optimal value of 20, the proposed approach outperforms the naïve approach by 2.1% on average. The largest improvement is achieved for the road surface class that the proposed approach shows a 3.2% increase in detection accuracy than the naïve one. The improvement seems to be insignificant because the classifiers are carefully tuned through extensive experiments as presented above. When the parameters are less optimized, e.g., when the number of trees is set to 1 as shown in the right half of Table 6, the proposed approach improves the classification accuracy by 6.4% on average, and 8% for the road surface category. Further studies into the results suggest an important reason for the above improvements lies in that the proposed temporal and spatial correlated classification approach can recover the classification errors made by the independent classifiers.

6.6.2 Compare with FHMMs and CHMMs. In this experiment, we compare the performance of the proposed classification model with two extensions of HMMs that are often used for activity recognition. The results are obtained by testing on the same data set as used in Sec. 6.3 with 10-fold-cross-validation.

The models selected for comparison are Factorial Hidden Markov Models (FHMMs) [10], and Coupled Hidden Markov Models (CHMMs) [4, 37]. As shown in Fig. 20, for FHMMs and CHMMs, the observation sequence is the sequence of window level feature vectors, and the four hidden state sequences are the label sequences for activities, road surfaces, shoes, and vehicles, respectively. During the training phase, since the observation sequence and the corresponding label sequences are already present, we train the models by direct counting [37]. When testing, given the observation sequence, we use the Viterbi algorithm to decode the label sequences. While both models model the temporal correlations, the difference between FHMMs and CHMMs is that the former is composed of four independent chains that only models the temporal correlation while the later adds cross-chain dependencies that can also model spatial correlations.

Table 7 lists the experiment results which suggest FHMMs and CHMMs also achieve reasonable detection accuracies of 71.5%, and 74.9%, respectively, considering the complexity of the task. CHMMs outperforms FHMMs by nearly 5% in average accuracy, which is possibly due to the contribution of the spatial correlations modeled

by the cross-chain dependencies. However, the results also suggest FHMMs and CHMMs are less accurate than the proposed Random Forest-based model. Based on our experiences during the experiments, we present the possible reasons as follows. First, it is difficult to precisely express the correlations between the hidden states and the observations represented by the 360-dimensional feature vector by simply using the emission probabilities. For Random Forest-based approaches, such correlation is first modeled by a collection of decision trees and then enhanced by the bagging technique, making the detection accurate even with the independent classifiers. Second, FHMMs and CHMMs are prone to serial failures which means the mistake made in one step can lead to mistakes in the following steps for a long period. Similar observations are also made during our previous work on multi-user activity recognition [13]. Additionally, the decoding process of the FHMMs and CHMMs is more time consuming than the proposed approach. Finally, as a discriminative model, the Random Forest classifier is expected to outperform generative models such as HMMs on classification tasks with fewer data provided [14].

In summary, the proposed model outperforms FHMMs and CHMMs on our data set. However, it is possible to improve their performance by tuning the model parameters or using more sophisticated models. Further discussions on these topics are out of the scope of this paper, which we leave for our future work.

6.6.3 Compare with Other Sensing Modalities. In this experiment, we compare the performance of the proposed system with audio- and inertial sensor-based approaches. Comparison data are collected by asking two male subjects to collect vibration, audio, and inertial sensor readings simultaneously during their normal daily activities for over two weeks. For audio-based approach, we use the same system parameters for feature extraction and detection. To conduct a fair comparison, a bluetooth microphone is attached to the same position of the vibration sensor on the other foot during audio data collection. For inertial sensor-based approach, we collected data by sampling the built-in sensors of the smartphone used for audio and vibration data collection. Data from two types of sensors—accelerometer and gyroscope—are collected and both time- and frequency-domain features [14] are computed.

Without external noise. Table 8 shows the comparison results which suggest that the proposed vibration-based approach achieves comparable accuracy with audio-based approach when no external noise is present. It also suggests both vibration- and audio-based approaches outperform inertial-sensor based approaches on road surface and shoe type detection. However, inertial sensor achieves comparable results on activity and vehicle type detection, which is consistent with the results of existing inertial sensor-based approaches [15]. Since our approach involves a smartphone for data collection and processing, inertial sensor-based approach can easily be integrated into our framework. The combined approach achieves the highest average detection accuracy of 97.2% as shown in the last column of Table 8.

With external noise. Though Table 8 suggests the average detection accuracy of the proposed approach is slightly lower than audio-based approach, the advantage of the proposed system lies in that it is privacy preserving and resistant to external sounds as discussed above. To further demonstrate this advantage, we conduct a simple controlled study by asking one subject to sit in a room wearing both the vibration and audio sensors with another subject walking around him. Detection results show that the proposed vibration sensor correctly detects the subject to be *idle* with 100% accuracy while audio-based approach mistakenly recognizes the subject to be *walking* in 66.7% of the time.

6.7 Discussion

We provide some discussions on the system's design and implementation in this section.

First, while the above experiment results suggest the system achieves high detection accuracy, we can further improve the system's performance by applying a post-processing step on the results. For example, we can use low pass filters or Markov models [21] to smooth the detection results. We omit this step in this work to show

the performance of the system without post-processing the results. We plan to add this function in our future system implementation to improve the system's performance.

Second, besides frame admission control and feature selection, we can further reduce the system's costs by introducing other low-cost sensors such as accelerometers or gyroscopes to provide more aggressive admission control policies. For example, users are not likely to change their shoes during running. As a result, we can turn off the vibration sensor to save energy when user's state detected by the low-cost sensors remains unchanged.

Finally, during data collection, we find the current vibration sensors are easy to break when performing strenuous exercises such as running. We plan to explore new ways of encapsulating the sensor to make it more robust while keeping its flexibility and sensitivity in our future work. We also plan to test different placement strategies to improve wearing experience and sensing performance.

7 POTENTIAL APPLICATIONS

There are many applications that can potentially be supported by the *SpiderWalk* system. In this section, we discuss some of them as follows.

Map generation and tracking. *SpiderWalk* can be used to generate maps with detailed road condition and surface information with participatory sensing applications supported by the *SpiderWalk* system. A *SpiderWalk* augmented tracking application can then accurately track a user's path even with inaccurate GPS positioning information, e.g., the user is walking along a brick sidewalk along the side of the main street. The tracking application can also warn the user on detecting the user has left the sidewalk and stepped into the main street for pedestrian safety protection [16].

We have built a sample application on top of *SpiderWalk* which can track the user's current transportation activity with rich circumstance information. It also keeps a log of user's activities and shoe preferences for persuasion and recommendation purposes. Based on this application, we are now laughing a crowdsourcing project to generate a detailed map with road surface information and analyze the spatial-temporal distribution of different activities on our campus.

Personal assistance. Personable preferences such as the most favorite shoes, preferred ways of traveling, and favorite exercises can be mined from the detection results. Moreover, it is even possible to infer use's gender, occupation, age, and habit through long-term activity and context analysis.

The rich circumstance information provided by *SpiderWalk* can make persuasive applications smarter. For example, the application can warn users for possible injuries when exercising in hazard environments or wearing improper shoes, or suggest the user to jog on roads instead of lawns to protect the environment.

Security and health. As an ultra-sensitive wireless vibration sensor, the applications of the proposed system is unlimited. It can be applied for person identification for securities. It can also in other application scenarios such as non-speech body sounds detection [26, 39].

We have also started a project aiming at capturing the abdominal surface vibrations by embedding the contact vibration sensor to the belt. The captured vibration signals are used to detect the bowel sounds produced by movements of the gastrointestinal (GI) tract, which could be indicators for diseases such as irritable bowel syndrome, GI bleeding, and other GI symptoms [28].

Due to its advantages such as device simplicity and low-cost, it is hopeful that *SpiderWalk* will be used in many real-world applications in the near future.

8 CONCLUSION

In this paper, we introduce the *SpiderWalk* system that achieves circumstance-aware transportation activity detection using a novel contact vibration sensor. By using a novel, flexible crack-resistance sensing material which is ultra-sensitive to vibrations, our system can capture the subtle vibration patterns produced by different

activities under different circumstances. Comparing to existing work on sensor-based transportation activity detection, our system can provide information on not only the activity but also its surrounding circumstances such as road surface, vehicle, and shoe types to meet the increasing demands of emerging applications as discussed above. By simulating a COTS electret microphone, the proposed contact vibration sensor can easily be integrated into COTS audio devices, making our system low-cost and ready-to-use in real-world applications. Moreover, because the contact vibration sensor is worn under the foot and is resistant to external sounds, our sensor outperforms electret microphone-based solutions on data assignment and privacy preserving issues. Experiments conducted on a real-world data set suggest our system achieves an average detection accuracy of 93.8%, showing the system's effectiveness. Resource consumption testing results show that the CPU, memory, and power overheads of the proposed system running on a smartphone is similar to existing audio- and GPS-based systems, suggesting the proposed system can provide rich circumstance information without increasing the system's resource consumption.

As a pilot study on using this novel contact vibration sensor for circumstance-aware transportation activity detection, there are many directions to follow in the future, including but not limited to: 1) more robust and optimized sensor design; 2) different placement strategies; 3) pre- and post-processing algorithms; 4) admission control and duty cycling approaches; and 5) other novel applications.

ACKNOWLEDGMENTS

The authors would like to thank all the participants involved in the study, and the anonymous editor and reviewers who helped in improving our manuscript. This work is supported by the NSFC No. 61690204, the National 973 Project No. 2015CB352202, NSFC No. 61502225, Australian Research Council Discovery Grant No. DP180103932, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Soleh Udin Al Ayubi. 2013. Model, Framework, and Platform of Health Persuasive Social Network. *Proquest Llc* (2013).

[2] Fahd Albinali, Stephen S Intille, William L Haskell, and Mary Rosenberger. 2010. Using wearable activity type detection to improve physical activity energy expenditure estimation. In *International Conference on Ubiquitous Computing (UbiComp'10)*. 311–320.

[3] Andreas Bloch, Robert Erdin, Sonja Meyer, Thomas Keller, and Alexandre de Spindler. 2015. Battery-Efficient Transportation Mode Detection on Mobile Devices. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, Vol. 1. IEEE, 185–190.

[4] Matthew Brand, Nuria Oliver, and Alex Pentland. 1997. Coupled hidden Markov models for complex action recognition. In *Computer vision and pattern recognition (CVPR'1997), 1997. proceedings., 1997 ieee computer society conference on*. IEEE, 994–999.

[5] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (2001), 5–32.

[6] Ke-Yu Chen, Rahul C Shah, Jonathan Huang, and Lama Nachman. 2017. Mago: Mode of Transport Inference Using the Hall-Effect Magnetic Sensor and Accelerometer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT'17)* 1, 2 (2017), 8.

[7] Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The WEKA Workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition* (2016).

[8] Jon Froehlich, Tawanna R Dillahunt, Predrag Klasnja, Jennifer Mankoff, Sunny Consolvo, Beverly Harrison, and James A Landay. 2009. UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In *Human Factors in Computing Systems (CHI'09)*.

[9] Petko Georgiev, Nicholas D Lane, Kiran K Rachuri, and Cecilia Mascolo. 2014. DSP.Ear: leveraging co-processor support for continuous audio sensing on smartphones. In *ACM Conference on Embedded Network Sensor Systems (SenSys'14)*.

[10] Zoubin Ghahramani and Michael I. Jordan. 1997. Factorial Hidden Markov Models. *Machine Learning* 1 (1997), 31.

[11] Hongmian Gong, Cynthia Chen, Evan Bialostozky, and Catherine T. Lawson. 2012. A GPS/GIS method for travel mode detection in New York City. *Computers Environment and Urban Systems (CEUS'12)* 36, 2 (2012), 131–139.

[12] Marta C Gonzalez, Cesar A Hidalgo, and Albertlaszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

[13] Tao Gu, Liang Wang, Hanhua Chen, Xianping Tao, and Jian Lu. 2011. Recognizing multiuser activities using wireless body sensor networks. *IEEE transactions on mobile computing (TMC'11)* 10, 11 (2011), 1618–1631.

[14] Tao Gu, Liang Wang, Zhanqing Wu, Xianping Tao, and Jian Lu. 2010. A Pattern Mining Approach to Sensor-Based Human Activity Recognition. *IEEE Transactions on Knowledge & Data Engineering (TKDE'10)* 23, 9 (2010), 1359–1372.

[15] Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2013. Accelerometer-based transportation mode detection on smartphones. In *International Conference on Embedded Networked Sensor Systems (SenSys'13)*.

[16] Shubham Jain, Carlo Borgiattino, Yanzhi Ren, Marco Gruteser, Yingying Chen, and Carla Fabiana Chiasserini. 2015. LookUp: Enabling Pedestrian Safety Services via Shoe Sensing. In *International Conference on Mobile Systems, Applications, and Services (MobiSys'15)*. 257–271.

[17] Daeshik Kang, Peter V Pikhitsa, Yong Whan Choi, C W Lee, Sung Soo Shin, Linfeng Piao, Byeonghak Park, Kahpyang Suh, Taeil Kim, and Mansoo Choi. 2014. Ultrasensitive mechanical crack-based sensor inspired by the spider sensory system. *Nature* 516, 7530 (2014), 222.

[18] Jeffrey P Koplan, David S Siscovick, and Gary M Goldbaum. 1985. The risks of exercise: a public health view of injuries and hazards. *Public Health Reports* 100, 2 (1985), 189.

[19] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. 2015. DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *International Conference on Ubiquitous Computing (UbiComp'15)*.

[20] Xiangyu Li, Xiao Zhang, Kongyang Chen, and Shengzhong Feng. 2014. Measurement and analysis of energy consumption on Android smartphones. In *IEEE International Conference on Information Science and Technology (ICIST'14)*.

[21] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2009. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *International Conference on Mobile Systems, Applications, and Services (MobiSys'09)*. 165–178.

[22] Denys J. C. Matthies, Thijs Roumen, Arjan Kuijper, and Bodo Urban. 2017. CapSoles: Who is Walking on What Kind of Floor?. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'17)*. ACM, New York, NY, USA, 9:1–9:14.

[23] Martin F McKinney and Jeroen Breebaart. 2003. Features for audio and music classification.. In *The International Society for Music Information Retrieval or International Symposium on Music Information Retrieval (ISMIR'03)*, Vol. 3. 151–158.

[24] Shahriar Nirjon, Robert F. Dickerson, Philip Asare, Qiang Li, Dezhi Hong, John A. Stankovic, Pan Hu, Guobin Shen, and Xiaofan Jiang. 2013. Auditeur: a mobile-cloud service platform for acoustic event detection on smartphones. In *International Conference on Mobile Systems, Applications, and Services (MobiSys'13)*.

[25] Thor Siiger Prentow, Henrik Blunck, Mikkel Baun Kjærgaard, and Allan Stisen. 2015. Towards Indoor Transportation Mode Detection Using Mobile Sensing. In *Mobile Computing, Applications, and Services (MobiCASE'15)*.

[26] Tauhidur Rahman, Alexander T. Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. BodyBeat: A mobile system for sensing non-speech body sounds. In *International conference on Mobile Systems, Applications, and Services (MobiSys'14)*. 2–13.

[27] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. 2010. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN'10)* 6, 2 (2010), 13.

[28] Annika Reintam, Pille Parm, Reet Kitus, Hartmut Kern, and Joel Starkopf. 2009. Gastrointestinal symptoms in intensive care patients. *Acta Anaesthesiologica Scandinavica* 53, 3 (2009), 318–324.

[29] Kartik Sankaran, Minhui Zhu, Xiang Fa Guo, Akkihebbal L. Ananda, Mun Choon Chan, and Li Shiuan Peh. 2014. Using mobile phone barometer for low-power transportation context detection. In *ACM Conference on Embedded Network Sensor Systems (SenSys'14)*. 191–205.

[30] Muhammad Awais Shafique and Eiji Hato. 2015. Use of acceleration data for transportation mode prediction. *Transportation* 42, 1 (2015), 163–188.

[31] Rahul C Shah, Chiehyih Wan, Hong Lu, and Lama Nachman. 2014. Classifying the mode of transportation on mobile phones using GIS information. In *International Conference on Ubiquitous Computing (UbiComp'14)*.

[32] Dongyoun Shin, Daniel G Aliaga, Bige Tuncer, Stefan Muller Arisona, Sungah Kim, Dani Zund, and Gerhard N Schmitt. 2015. Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems (CEUS'15)* 53 (2015), 76–86.

[33] L. Shu, T. Hua, Y. Wang, Li Q Qiao, D. D. Feng, and X. Tao. 2010. In-shoe plantar pressure measurement and analysis system based on fabric pressure sensing array. *IEEE Transactions on Information Technology in Biomedicine (TITB'10)* 14, 3 (2010), 767–75.

[34] Bernhard Slawik. 2014. ShoeSoleSense for Peripheral Interaction. In *Proceedings of the Workshop on Peripheral Interaction: Shaping the Research and Design Space at CHI 2014*. ACM.

[35] Timothy Sohn, Alex Varshavsky, Anthony LaMarca, Mike Y. Chen, Tanzeem Choudhury, Ian Smith, Sunny Consolvo, Jeffrey Hightower, William G. Griswold, and Eyal de Lara. 2006. Mobility Detection Using Everyday GSM Traces. In *Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp'06)*. 212–224.

[36] M Torkki, A Malmivaara, N Reivonen, S Seitsalo, P Laippalo, and V Hoikka. 2002. Individually fitted sports shoes for overuse injuries among newspaper carriers. *Scandinavian Journal of Work Environment & Health (SJWEH'02)* 28, 3 (2002), 176–183.

[37] Liang Wang, Tao Gu, Xianping Tao, Hanhua Chen, and Jian Lu. 2011. Recognizing multi-user activities using wearable sensors in a smart home. *Pervasive and Mobile Computing (PMC'11)* 7, 3 (2011), 287–298.
[38] Peter Widhalm, Philippe Nitsche, and Norbert Brandie. 2012. Transport mode detection with realistic Smartphone sensor data. In *International Conference on Pattern Recognition (ICPR'12)*.
[39] Koji Yatani and Khai N. Truong. 2012. BodyScope: a wearable acoustic sensor for activity recognition. *International Conference on Ubiquitous Computing (UbiComp'12)* (2012), 341–350.
[40] Zelun Zhang and Stefan Poslad. 2012. Fine-Grained Transportation Mode Recognition Using Mobile Phones and Foot Force Sensors. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services (Mobiquitous'12)*. 103–114.
[41] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei Ying Ma. 2010. Understanding transportation modes based on GPS data for web applications. *Acm Transactions on the Web (TWEB'10)* 4, 1 (2010), 495–507.

Received May 2017; revised November 2017; revised January 2018; accepted January 2018