SOUND-BASED TRANSPORTATION MODE RECOGNITION WITH SMARTPHONES

Lin Wang[1,2], *Daniel Roggen*[1]

[1]Wearable Technologies Lab, Sensor Technology Research Centre, University of Sussex, UK
[2]Centre for Intelligent Sensing, Queen Mary University of London, UK
lin.wang@qmul.ac.uk; daniel.roggen@ieee.org

ABSTRACT

Smartphone-based identification of the mode of transportation of the user is important for context-aware services. We investigate the feasibility of recognizing the 8 most common modes of locomotion and transportation from the sound recorded by a smartphone carried by the user. We propose a convolutional neural network based recognition pipeline, which operates on the short-time Fourier transform (STFT) spectrogram of the sound in the log domain. Experiment with the Sussex-Huawei locomotion-transportation (SHL) dataset on 366 hours of data shows promising results where the proposed pipeline can recognize the activities Still, Walk, Run, Bike, Car, Bus, Train and Subway with a global accuracy of 86.6%, which is 23% higher than classical machine learning pipelines. It is shown that sound is particularly useful for distinguishing between various vehicle activities (e.g. Car vs Bus, Train vs Subway). This discriminablity is complementary to the widely used motion sensors, which are poor at distinguish between rail and road transport.

Index Terms— Computational auditory scene analysis; context-awareness; convolutional neural network; sound event classification; transportation mode recognition

1. INTRODUCTION

The mode of transportation of smartphone users is an important type of contextual information that denotes users' mobility status during travel, such as walking, cycling or driving [1]. Analyzing such multimodal data enables context-aware applications in fields such as intelligent service adaptation, individual environmental impact monitoring, human-centered activity monitoring [2, 3].

In recent years, there have been numerous studies showing how to recognize transportation modes from motion (accelerometer, gyroscope and magnetometer) and global positioning system (GPS) sensors that are embedded in smartphone devices [4–10]. These approaches typically employ classical machine learning and deep learning pipelines to infer the mode of transportation from the sensor data. A majority of the research effort has been placed on motion sensors, as they are comparatively much less energy demanding compared to continuous GPS sensing, and they provide richer information about the movement (e.g. vibration and change in orientation) of the mobile device in comparison to GPS. The state of the art in motion-based transportation recognition performance was established in the SHL recognition challenge 2018 through an open international competition among 20 research teams [11, 12]. The outcomes reveal that approaches based on motion sensors struggle distinguishing between distinct transportation modes of similar classes: for example between train and subway (rail transport) or between bus and car (road transport). Recently, vision-based transportation recognition has also been reported [16].

Sound is an important modality that is available in smartphone devices and has been increasingly used to infer the context of ambient environment with multiple advantages. First, it is a complementary modality to motion. It may outperform motion-based context recognition for some classes, or help in their disambiguation through data fusion. Second, it is capable of providing broader contextual information. For instance, the recent challenge on detection and classification of acoustic scenes and events (DCASE) aims to classify various sound events in domestic and wild environments [13, 14]. Finally, manufacturers of mobile processors are placing significant emphasis on including hardware acceleration for sound processing pipelines to enable always-on sound-based interaction at low energy cost (e.g. to detect "ok google" on Android devices). So far, only few work has been placed on sound-based transportation mode recognition [19, 20]. An in-depth analysis on the recognition performance for a large variety of transportation activities has not been reported yet. There are mainly two challenges that hinder the progress in this field. First, most public locomotion and transportation datasets contain only motion and GPS sensor data and do not have sound data available [8, 15]. Second, smartphone recordings in real-life environments contain overlapped sound from the travelling vehicle, human, and the environment, which makes the recognition of transportation activities a challenging task.

In this paper we aim to answer a research question that has not been well addressed: *Can sound be used to detect the transportation mode of the user effectively?* We use the state-of-the-art Sussex-Huawei locomotion-transportation (SHL) dataset that contains multimodal data recorded by smartphone sensors, including from microphones[1]. We present different pipelines, including classical machine learning and emerging deep learning approaches, to recognize eight transportation activities (Still, Walk, Run, Bike, Car, Bus, Train, Subway) from the sound recorded by smartphones. We evaluate the recognition performance using the same train/test partitioning and data segmentation scheme as the SHL recognition challenge [11]. Recognition using convolutional neural networks shows promising results on the evaluation dataset, and demonstrates complementary performance between sound and motion sensors.

2. DATASET

The SHL dataset is a major outcome of a large-scale longitudinal data collection campaign, which collected 2812 hours of labeled data over a period of 7 months corresponding to 17,562 km in the south-east of the UK including London [17, 18]. The SHL dataset

[1]http://www.shl-dataset.org/

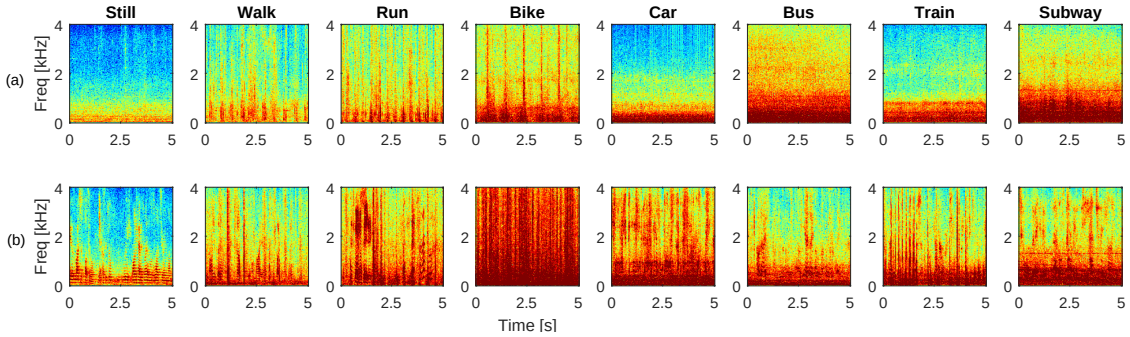

Fig. 2. Spectrogram of sound clips (5 seconds) for each transportation activity. (a) Clean sound. (b) Noisy sound with environmental noise.

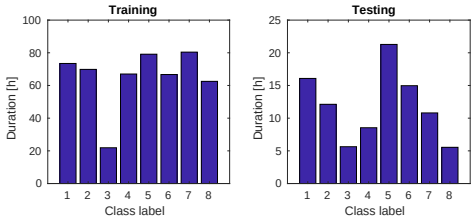

Fig. 1. The duration of each class activity in the training and the testing dataset. The 8 class activities are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

was recorded by three participants engaging in eight transportation and locomotion activities in real-life settings: Still, Walk, Run, Bike, Car, Bus, Train and Subway. Each participant carried four smartphones at four body positions simultaneously: hand, torso, hips and bag. Each smartphone logged the data of 16 sensors that are available in the device, including motion sensors, GPS, ambient pressure sensor, microphone, etc. The dataset is one of the biggest public dataset in the research community and the first dataset that contains sound modality. The dataset served as a main source feasting the recent SHL challenge: a competition on motion sensor-based transportation activity recognition [11, 12].

The sound recording in the SHL dataset enables us to investigate the feasibility of sound-based transportation mode recognition. For ease of comparison, we use exactly the same data as in the SHL challenge and abide by the same definition of training and testing data partitioning. Specifically, we use the sound recorded by the first participant with hand smartphone during 82 days (5-8 hours per day), which is partitioned in 62 days (271 hours) for training and 20 days (95 hours) for testing. Fig. 1 depicts the duration of each class activity in the training and testing datasets. The sound was originally recorded at a sampling rate of 16 kHz, and downsampled to 8 kHz.

Fig. 2 compares the short-time Fourier transform (STFT) spectrogram of the sound recorded during the 8 transportation activities, with and without environmental noise. In Fig. 2(a), the clean sound of each activity (without environmental noise) tends to show different spectrogram patterns. For instance, the activities Still, Car, Bus, Train and Subway tend to present different energy distribution in the low and high frequencies, while the activities Walk, Run and Bike tend to present different cyclic behaviour. This observation grounds the feasibility of sound-based transportation mode recognition. In practice, the clean sound of each transportation activity is usually overlapped with environmental noise, such as

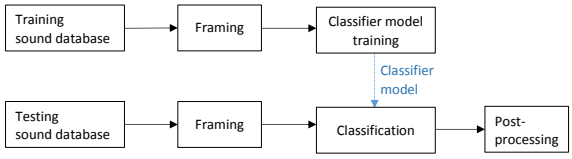

Fig. 3. Pipeline for sound-based transportation mode recognition.

wind, friction, human speech, and other sound events nearby, as shown in Fig. 2(b). These environmental noises are typically much stronger than the clean transportation sound. This significantly increases the challenges when recognizing transportation activities.

3. RECOGNITION PIPELINE

Fig. 3 depicts the processing pipeline for predicting transportation mode from sound, which is segmented into 5-second frames (40000 samples) in the training and testing datasets. The sound frames in the training dataset is used to train a classifier model, which is used to infer the mode of transportation in each sound frame in the testing dataset. A post-processing step follows to improve the recognition result by smoothing decisions across consecutive frames.

The size of raw sound data is 58 GB and 20 GB in the training and testing datasets, respectively. This large amount of data imposes additional challenges in the training stage. To minimize memory needs during training, we subsample the frames in the training data with a ratio of $1/4$, i.e. using a sliding window of 5 seconds long and jump size 20 seconds. For testing data, we do not conduct subsampling, i.e. using a sliding window of 5 seconds long and jump size 5 seconds. Finally, we have 52,091 training frames and 55,818 testing frames.

We consider two different types of classifiers: the classical classifier and the convolutional deep neural network. The former one performs feature computation and classification independently while the latter one learns the features and the classifier (deep neural network) simultaneously from the training data.

3.1. Classical machine learning

We extract two types of basic features, which are suggested in [21], in each sound frame: zero-crossing rate and mel-frequency cepstral coefficients (MFCC). The former is a very simple yet useful feature, whereas the latter is ubiquitous in speech processing and analyzing harmonic content. In each 5-second frame, the MFCC and the zero-

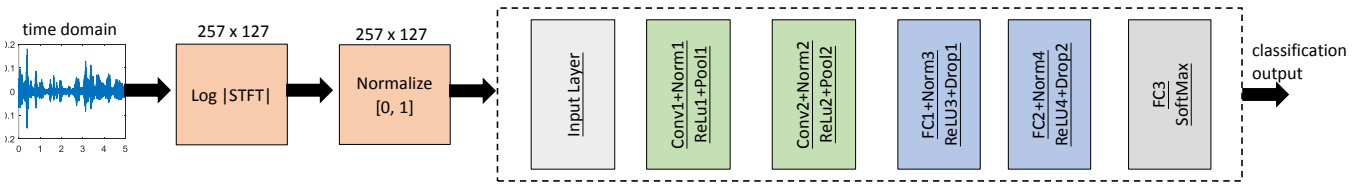

Fig. 4. Sound-based transportation mode recognition using a convolutional deep neural network.

crossing rate are computed with a sliding window of size 32 ms (256 samples) and half overlap. We compute the first 13 MFCC coefficients and the zero-crossing rate in each 32 ms subframe, and then summarize their mean and standard deviation across the 5-second frame. This leads to 28 features (14 mean and 14 std) in each sound frame.

We feed the feature vectors as input to five types of classifiers: naive Bayesian (NB), k-nearest neighbours (KNN), decision tree (DT), random forest (RF) and support vector machine (SVM).

3.2. Deep learning

Fig. 4 depicts a pipeline for sound-based transportation mode recognition using a convolutional neural network (CNN). CNN is a typical multi-layer neural network first proposed for computer vision problems and it is very suitable for image-related applications in machine learning. In recent years, the application of CNN to environment sound classification has been widely reported [22–25]. We compute STFT spectrogram in each sound frame and then feed it as an image input to the CNN classifier.

3.2.1. STFT Input

The STFT spectrogram is computed in each 5-second frame with a sliding window of length 64 ms (512 samples) and half overlap. Let's represent the STFT in a frame as $S(k, l)$, where k and l denote the frequency and the STFT subframe indices, respectively. In this way, the size of the spectrogram in one sound frame is 257×157.

To reduce the dynamic range of the data, we compute the log spectrogram as $A(k, l) = \log_{10} |S(k, l)|$, where $|\cdot|$ denotes the absolute value. We then normalize the data to the range of $[0, 1]$ as $I(k, l) = \frac{A(k,l) - A_{min}}{A_{max} - A_{min}}$, where A_{max} and A_{min} denote the maximum and the minimum values in the log spectrogram $A(:, :)$ throughout the training dataset.

3.2.2. CNN architecture

A convolutional neural network typically consists of a number of neural layers stacked together in a deep architecture: an input layer, several CNN and fully-connected neural network (FCNN) blocks, and an output decision block.

The input layer receives and stores the original spectrogram image I. Each CNN block sequentially consists of a convolutional layer, a batch normalization (Norm) layer, a nonlinear (ReLu) layer and a pooling layer. The *convolutional* layer puts the input spectrogram image through a set of convolutional filters, each of which activates certain features from the input. The convolutional layer is defined by the number of filters, the size of the filter, and the step size (stride) when traversing the input. The rectified linear unit (*ReLU*) layer allows for faster and more effective training by mapping negative values to zero and maintaining positive values, e.g. by using the activation function $f(x) = \max(0, x)$. In this way, only the activated features are carried forward into the next layer. The *batch normalization* layer normalizes the filtering output across a mini-batch, in order to speed up training of the neural network and to reduce the sensitivity to network initialization. The *pooling* layer simplifies the output by performing nonlinear downsampling, reducing the number of parameters that the network needs to learn. We employ the max-pooling method, where the input is divided into rectangles (pool) and the largest is taken from among all sub-pieces. The pooling layer is defined by the size of the pool and the stride when traversing the input.

Each FCNN block sequentially consists of a fully-connected (FC) layer, a batch normalization (Norm) layer, a nonlinear (ReLu) layer and a dropout layer. The *FC* layer consists of a number of neurons which are connected to all the neurons in the previous layer. The FC layer is defined by the number of neurons. The *batch normalization* layer normalizes the neuron outputs while the *ReLu* layer performs non-linear activation. The *dropout* layer randomly sets input elements to zero with a given probability in order to prevent overfitting [26].

The decision block consists of a FC layer, a nonlinear (Softmax) layer which outputs the classification result. The FC layer contains a number of neurons equalling to the number of decision labels.

Fig. 4 shows the proposed convolutional neural network architecture which consists of one input layer, two CNN blocks, two FCNN blocks and one output block. The parameters of neural network are given in Table 1. Due to the large amount of data, we employ a mini-batch processing scheme which updates the weights of the neural network per subset of training samples. We set the maximum number of epoches as 30 during training.

Table 1. Parameters of the CNN architecture.

Input layer	size: (257, 127)
Conv1/Conv2	number: 32; size: (5,5); stride: (1,1); padding: (0,0)
Pool1/Pool2	max pooling: (2,2); stride: (1,1); padding: (0,0)
FC1/FC2	nodes: 300
Drop1/Drop2	50%
FC3	nodes: 8
Norm1-4	mini-batch: 150

3.3. Post-processing

The recognition pipeline makes a decision per frame (5 seconds). Since the transportation mode of a user typically continues for a certain period and there is a strong correlation between neighbouring frames, we reasonably assume that the transportation mode remains the same in a short period (segment). Based on this assumption, we employ a majority voting scheme to further improve the recognition performance. Specifically, for a short segments with F sound frames, the decision is unified as the class activity that occurs mostly in these F frames [12]. To obtain consistent results with the SHL challenge, which chopped the data into 1-minute blocks, we also set

Table 2. Sound-based transportation mode recognition performance.

Classifier	Performance [%]		Processing time [s]	
	Accuracy	F1 score	Training	Testing
NB	54.6	51.5	0.31	0.1
DT	54.8	50.9	0.93	0.03
RF	**62.8**	58.3	5.0	0.6
KNN	59.4	57.4	0.04	17.1
SVM	58.8	53.0	223.6	0.18
CNN	80.6	77.9	39848	70.3
CNN+PP	**86.6**	85.6		

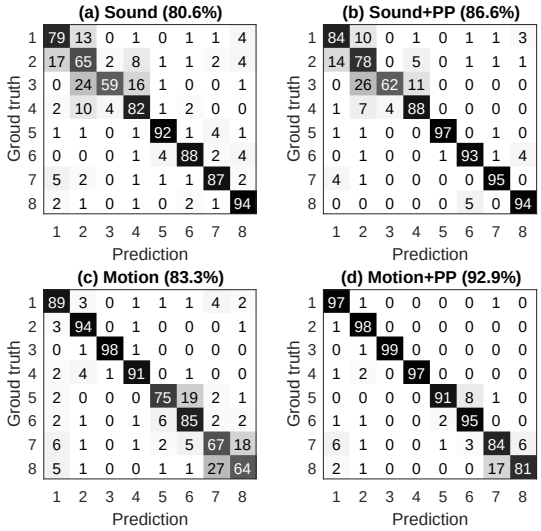

Fig. 5. Confusion matrices achieved with different modalities. (a)(b): using sound and post-processing. (c)(d): using motion sensors and post-processing. The 8 classes: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

the length of the smoothing segment to be 1 minute [12].

4. EVALUATION RESULTS

We use a computer equipped with an Intel i7-4770 4-core CPU @ 3.40 GHz with 32 GB memory, and a GeForce GTX 1080 Ti GPU with 3584 CUDA cores @ 1.58 GHz and 11 GB memory. The code is written with Matlab 2018a, calling functions from the Machine Learning Toolbox and the Deep Learning Toolbox. The recognition performance of the classifier is evaluated on the testing data in terms of global accuracy and F1 score [11].

Table 2 compares the results achieved by various classifiers in terms of performance and processing time. For the same classifier, the accuracy and the F1 score do not show significant difference. We thus only compare the global accuracy.

All the classical classifiers perform poorly on the testing data with RF achieving the highest accuracy of 62.8%. The deep-learning CNN pipeline (80.6%) outperforms classical classifiers remarkably over classical classifiers. In this paper we employed a very simple feature extraction scheme for the classical classifier, potentially leading to the poor performance. In addition, the hyper-parameters of these classifiers are not optimized. Deep learning does not rely on a pre-defined feature extraction stage. Instead, the convolutional layers, which act as feature extractors, are optimised as part of the learning process, which explains the significantly better performance when sufficient training data is available. The processing time, however, is more significant: RF takes 5 seconds for training and 0.6 seconds for testing, while CNN takes 39,848 seconds for training and 70 seconds for testing despite using GPU acceleration. In addition, the simple post-processing scheme (86.6%) can further improve recognition performance effectively by exploiting the temporal correlation between consecutive frames.

Fig. 5 compares the confusion matrices achieved by using sound alone and by using motion sensors (accelerometer, gyroscope and magnetometer) alone. For sound, we use the result achieved by CNN and CNN+PP. For motion sensors, we use the results that were reported in Fig. 5 in [12], which used CNN on the frequency-domain raw data and post-processing. Note that the two groups of results are comparable as they use exactly the same train/test partitioning and data segmentation.

The comparison leads to several interesting observations. Sound is better at classifying the vehicle activities (Car, Bus, Train and Subway) than motion sensors. This is because each vehicle transportation typically emits unique sound that distinguishes itself from other activities, but presents similar motion patterns. Motion sensor is better at classifying pedestrian activities (Still, Walk, Run, Bike) than sound. This is because pedestrian and biking activities require strong user engagement, but emit sound which is much weaker than environmental noise. Overall, the recognition results using the sound and using the motion sensors are truly complementary. The implies that the combination of the two can potentially lead to better recognition result.

5. CONCLUSION

We investigate the possibility of using sound to recognize transportation mode and propose a deep-learning CNN network operating on the STFT log spectrum of the sound. Experimental results validate the feasibility of sound-based transportation mode recognition, and demonstrate that CNN outperforms classical classifiers with the set of features we selected here. The classification result based on sound is complementary to the one based on motion sensors, where the former one is good at recognizing vehicle activities and the latter is good at recognition pedestrian and biking activities.

As one of the first works that systematically investigate sound-based transportation mode recognition, the paper foresees several future directions of research. First, the recognition performance using sound can be improved by optimizing the CNN architecture, and by combining with other modalities such as motion and GPS sensors. Second, sound tends to work more robustly in case of user and sensor placement variation than motion sensors. It would be interesting to investigate the recognition performance using sound in the full SHL dataset, which includes various users and smartphone positioning. Finally, the audio dataset is currently not publicly available, for ethical approval and privacy protection reasons. The sound recordings have never been listened to by the researchers, and all the process has been automated looking only at the 8 indicated classes. Future work will investigate means to release this audio dataset while preserving the privacy of users.

Acknowledgement: This work was supported by the HUAWEI Technologies within the project "Activity Sensing Technologies for Mobile Users". This work was also supported by the Institute of Coding, which is supported by the Office for Students (OfS) and the Higher Education Funding Council for Wales (HEFCW). We thank NVidia for GPU donation.

6. REFERENCES

[1] J. Engelbrecht, M. J. Booysen, G. J. van Rooyen, and F. J. Bruwer, "Survey of smartphone-based sensing in vehicles for intelligent transportation system applications," *IET Intelligent Transport Systems*, vol. 9, no. 10, pp. 924-935, 2015.

[2] G. Castignani, T. Derrmann, R. Frank, and T. Engel, "Driver behavior profiling using smartphones: A low-cost platform for driver monitoring," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 1, pp. 91-102, 2015.

[3] Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Computing*, vol. 16, no. 4, pp. 62-74, 2017.

[4] P. Siirtola and J. Roning, "Recognizing human activities user independently on smartphones based on accelerometer data," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 1, no. 5, pp. 38-45, 2012.

[5] T. Feng and H. J. Timmermans, "Transportation mode recognition using GPS and accelerometer data," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 118-130, 2013.

[6] S. Hemminki, P. Nurmi, and S. Tarkoma, "Accelerometer-based transportation mode detection on smartphones," in *Proc. ACM Conf. Embedded Networked Sensor Systems*, Roma, Italy, 2013, pp. 1-14.

[7] H. Xia, Y. Qiao, J. Jian, and Y. Chang, "Using smart phone sensors to detect transportation modes," *Sensors*, vol. 14, no. 11, pp. 20843-20865, 2014.

[8] M. C. Yu, T. Yu, S. C. Wang, C. J. Lin, and E. Y. Chang, "Big data small footprint: The design of a low-power classifier for detecting transportation modes," in *Proc. Very Large Data Base Endowment*, Hangzhou, China, 2014, pp. 1429-1440.

[9] X. Su, H. Caceres, H. Tong, and Q. He, "Online travel mode identification using smartphones with battery saving considerations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2921-2934, 2016.

[10] S. H. Fang, Y. X. Fei, Z. Xu, and Y. Tsao, "Learning transportation modes from smartphone sensors based on deep neural network," *IEEE Sensors Journal*, vol. 17, no. 18, pp. 6111-6118, 2017.

[11] L. Wang, H. Gjoreski, K. Murao, T. Okita, and D. Roggen, "Summary of the Sussex-Huawei locomotion-transportation recognition challenge," in *Proc. 2018 ACM Int. Joint Conf. 2018 Int. Sym. Pervasive Ubiquitous Computing Wearable Computers*, Singapore, 2018, pp. 1521-1530.

[12] L. Wang, H. Gjoreski, M. Cilliberto, S. Mekki, S. Valentin, and D. Roggen, "Benchmarking the SHL recognition challenge with classical and deep-learning pipelines," *Proc. 2018 ACM Int. Joint Conf. 2018 Int. Sym. Pervasive Ubiquitous Computing Wearable Computers*, Singapore, 2018, pp. 1626-1635.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: an overview of dcase 2017 challenge entries," in *Proc. Int. Workshop Acoust. Sig. Enhancement*, Tokyo, Japan, 2017, pp. 1-5.

[14] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 2, pp. 379-393, 2018.

[15] Y. Zheng, X. Xie, and W. Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data Engineering Bulletin*, vol. 33, no. 2, pp. 32-39, 2010.

[16] S. Richoz, M. Ciliberto, L. Wang, P. Birch, H Gjoreski, A. Perez-Uribe, D. Roggen, "Human and machine recognition of transportation modes from body-worn camera images," in *Proc. Joint 8th Int. Conf. Informatics, Electronics & Vision and 3rd Int. Conf. Imaging, Vision & Pattern Recognition*, Washington, USA, 2019, pp. 1-6.

[17] H. Gjoreski, M. Cilliberto, L. Wang, F. J. O. Morales, S. Mekki, S. Valentin, and D. Roggen, "The University of Sussex-Huawei locomotion and transportation dataset for multimodal analytics with mobile devices," *IEEE Access*, vol. 6, pp. 42592-42604, 2018.

[18] L. Wang, H. Gjoreski, M. Cilliberto, S. Mekki, S. Valentin, and D. Roggen, "Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset," *IEEE Access*, vol. 7, pp. 10870-10891, 2019.

[19] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The Jigsaw continuous sensing engine for mobile phone applications," in *Proc. ACM Conf. Embedded Networked Sensor Systems*, Zurich, Switzerland, 2010, pp. 71-84.

[20] S. Lee, J. Lee, and K. Lee, "VehicleSense: A reliable sound-based transportation mode recognition system for smartphones," in *Proc. IEEE 18th Int. Symp. World Wireless, Mobile Multimedia Networks*, Macau, China, 2017, pp. 1-9.

[21] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Int. Conf. Multimedia*, Brisbane, Australia, 2015, pp. 1015-1018.

[22] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130-133, 2011.

[23] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE Int. Workshop Machine Learning for Signal Processing*, Boston, USA, 2015, pp. 1-6.

[24] S. Hershey, S. Chaudhuri, D. P. Ellis, et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, New Orleans, USA, 2017, pp. 131-135.

[25] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network", *Neurocomputing*, vol. 272, pp. 505-512, 2018.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.