Embedded Identification of Surface Based on Multirate Sensor Fusion With Deep Neural Network

Semin Ryu, *Member, IEEE*, and Seung-Chan Kim, *Member, IEEE*

***Abstract*—In this letter, we propose a multivariate time-series classification system that fuses multirate sensor measurements within the latent space of a deep neural network. In our network, the system identifies the surface category based on audio and inertial measurements generated from the surface impact, each of which has a different sampling rate and resolution in nature. We investigate the feasibility of categorizing ten different everyday surfaces using a proposed convolutional neural network, which is trained in an end-to-end manner. To validate our approach, we developed an embedded system and collected 60 000 data samples under a variety of conditions. The experimental results obtained exhibit a test accuracy for a blind test dataset of 93%, taking less than 300 ms for end-to-end classification in an embedded machine environment. We conclude this letter with a discussion of the results and future direction of research.**

***Index Terms*—Deep learning, latent space, multirate measurements, multivariate measurement, sensor fusion, time-series classification.**

I. Introduction

An INTELLIGENT system that incorporates multiple sensors for time-series classification purposes often requires a sophisticated multisensor fusion method in that measurements from each sensor are generally not sampled at the same rate. Although multirate sensor measurements can be fused using a conventional approach (e.g., a direct weighted fusion), such methods often result in a limited applicability owing to their simplicity [1]. Although a fusion can be achieved by modeling the multirate sensor system [2], this type of approach requires additional information on the complex system dynamics.

Taking advantage of recent deep learning capabilities, a recent study proposed a temporal binding approach that classifies audio–visual information based on an efficient multimodal fusion [3]. In this letter, a set of temporal information, including the RGB flow and audio, is efficiently fused in a latent space of a convolutional neural network (CNN) such that all modalities are trained simultaneously. To the best of our knowledge, few studies have addressed the time-series classification of multirate multivariate sensor measurements that include heterogeneous time-series measurements, such as accelerations and audio recordings.

Herein, we propose an intelligent system that identifies various surfaces by hitting them autonomously and interpreting the resulting multivariate measurements. We used a custom-built hardware setup comprising a solenoid actuator, microphone, triaxial accelerometer, microcontroller, and an embedded machine. The captured multirate sensor streams were analyzed using a machine learning pipeline that fuses multisensor measurements within the latent space. We conducted a series of experiments, including those on the test accuracy and inference time, to assess the feasibility of the proposed approach. The results demonstrate that our system can successfully classify various surface categories on an embedded machine.

II. Related Studies

Several research groups have attempted to classify objects based on their physical contact. BeatIt [4] used a smartwatch to study the categorization of objects based on the sound generated when a user knocked on them. Cho *et al.* [5] enabled smartphones to identify underlying objects by generating a vibration and then interpreting the resulting linear accelerations. These methods demonstrate the feasibility of object recognition with a single sensor; however, the test accuracy of approximately 80% that was achieved needs to be further improved. One way to achieve a better performance would be through the fusion of heterogeneous sensors. In general, multimodal sensor fusion causes difficulties in the analysis mainly owing to intrinsic differences among the sensor data. Owing to the different data types and sample rates applied, each modality is characterized by distinctive statistical properties, representations, and correlation structures. Cross-modality relationships are highly nonlinear and difficult to determine even by hand [6]. Using a samrtphone, Knocker [7] tried to identify various daily objects using multisensor data generated when knocking on them. They adopted a support vector machine (SVM) classifier, and the classification latency (or inference time) was measured at approximately 229 ms. Although their test accuracy for real-world data, i.e., when using a blind test set, reached approximately 83%, this rate can be further improved by

Manuscript received March 20, 2020; accepted May 16, 2020. Date of publication May 22, 2020; date of current version May 27, 2021. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korea Government (MSIT) under Grant 2019-0-00050 (Brain-Inspired AI Framework and Cognitive Convergence Research and Development Pipeline-Based Health Behavior Monitoring, Diagnosis, and Prescription Technology), and in part by the 2018 Hallym University Research Fund under Grant HRF-201803-001. This manuscript was recommended for publication by Y. Chen. *(Corresponding author: Seung-Chan Kim.)*

The authors are with the Intelligent Robotics Laboratory, Hallym University, Chuncheon 24252, South Korea, and also with the Hallym Institute for Data Science and Artificial Intelligence, Hallym University, Chuncheon 24252, South Korea (e-mail: dalek@glab.hallym.ac.kr).

Digital Object Identifier 10.1109/LES.2020.2996758

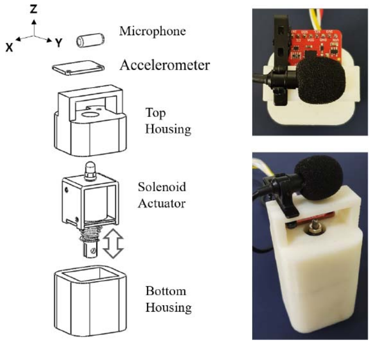


Fig. 1. Custom-built prototype used to knock on underlying surfaces and measure the resulting multisensor signals. The coordinate system shown in the upper left part denotes the orientation of the accelerations.

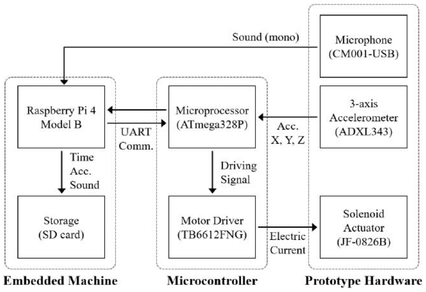


Fig. 2. Entire block diagram of the proposed system. The prototype is controlled by an embedded machine through a microcontroller.

applying a deep learning approach instead of a shallow learning method. Meanwhile, Radu *et al.* [8] compared shallow and deep learning approaches for several public multisensor datasets. The results indicate that, in numerous cases, deep learning outperforms shallow learning techniques.

Contrary to existing studies, our method achieves a high accuracy by virtue of fusing multirate sensor data using a deep learning architecture. By designing a compact network structure, the inference time is sufficiently short even on an embedded environment. Furthermore, the proposed system does not require user intervention (e.g., a knocking action), and can be incorporated into various smart devices, such as an artificial intelligence (AI) speaker.

III. Proposed Approach

A. Dataset

A custom hardware setup was constructed to collect the dataset. Fig. 1 shows the prototype hardware used to generate the physical impact and measure the resulting audio and acceleration signals, and Fig. 2 shows the overall structure of the constructed setup. The solenoid actuator (JF-0826B, Yeuqing Gangbei Electric) inside the housing impulsively contacts the surface of interest through a push–pull operation. The microphone (CM001-USB, Comsonic) and accelerometer (ADXL343, Analog Devices) mounted on top of the prototype measure the audio and triaxial acceleration signals. As shown in Fig. 2, the embedded machine (Raspberry Pi 4 Model B) controls the overall system through a microcontroller (ATmega328P, Arduino Nano). The resulting audio and acceleration signals are recorded while the solenoid actuator hits the surface.

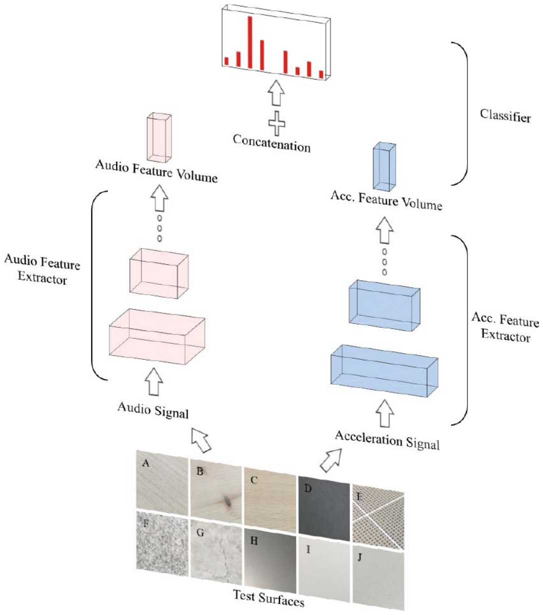


Fig. 3. Proposed late fusion CNN architecture. The network utilizes a split architecture with separate branches for each of two modalities, which are then merged into the cross-sensor layer.

We considered the following ten surfaces found in our daily surroundings: 1) a softwood table; 2) a synthetic wood table; 3) a plywood table; 4) a polyurethane chair; 5) a rubber cutting mat; 6) a granite tiled floor; 7) a porcelain tiled floor; 8) a metal plate; 9) an acrylonitrile butadiene styrene (ABS) plastic table; and 10) a laminate table. For each category, 100 data samples were collected at each of ten randomly chosen locations of the surfaces (including near the center, edges, and corners for the tables) with three different impact intensities. The intensity was controlled in three steps by restricting the stroke of the metal slug using rubber damping stoppers. The entire collection process was repeated one more time for different objects. Hence, in total, 6000 data samples (100 samples $\times$ 10 locations $\times$ 3 impact intensities $\times$ 2 objects) were collected for each surface. During the data collection, background noise (auditory and vibratory), such as music and machinery noise were randomly played. Each sample was approximately 2.4 s long, and the sampling frequencies were 44.1 kHz and 250 Hz for the audio and acceleration signals, respectively. The trimmed signal (0.4 s), which includes the impact motion, was used as a raw input signal. We separated the dataset into two independent sets. 70% of the dataset was used as a training set and the remainder was applied for the validation set.

B. Machine Learning Pipeline

In Fig. 3, a schematic of the proposed approach toward multimodal learning using a feature concatenation is shown. First, the hardware is used to knock on the surface, and the resulting audio and acceleration signals are captured. Second, the features from each modality are extracted individually.

TABLE I
FEATURES EXPLORED IN THIS LETTER

Modality	Extracted features
Audio	MFCCs, zero crossing rate Spectral rolloff, spectral centroid Spectral contrast, spectral bandwidth
Acceleration	Arithmetic mean (average), median Minimum, maximum, ratio of max and min Standard deviation, sample skewness Minimum of absolute value Maximum of absolute value Arithmetic mean of absolute value Standard deviation of absolute value

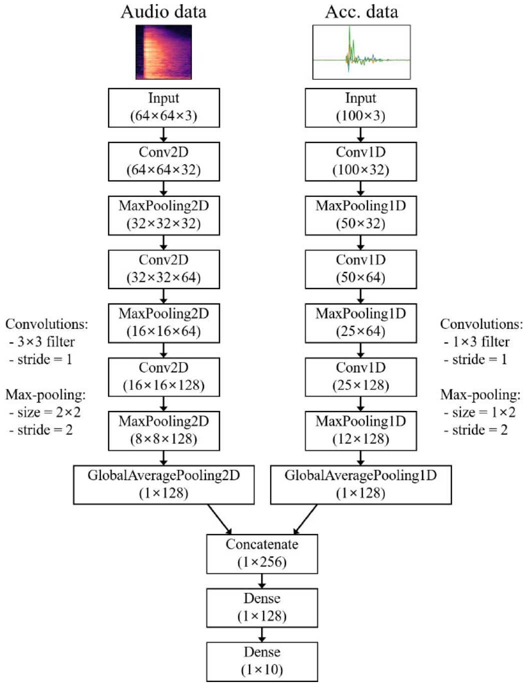


Fig. 4. Proposed network architecture of multimodal deep learning approach.

Finally, the extracted features are combined into a single feature vector presented to a classifier for identification across all features. We can implement this strategy in two ways: through a shallow learning approach, i.e., using a random forest (RF), and by applying a deep learning approach, i.e., using a CNN.

C. Shallow Learning Approach

As a baseline, we employed an RF classifier owing to its robustness against an overfitting [9]. For the audio signals, mel-frequency cepstral coefficients (MFCCs) and spectral features were calculated as a feature set. For the acceleration signals, we used a statistical feature set. These feature sets were determined after exploring different feature sets, including the spectral features (magnitude spectrum, log magnitude spectrum, etc.) However, they only slightly improved the test accuracy at the expense of further computational burden or worsened the accuracy. In Table I, all features used in this letter are summarized. The RF classifier was implemented in Python 3.6 using scikit-learn (scikit-learn.org).

TABLE II
EXPERIMENTAL RESULTS: TEST ACCURACY FOR THE BLIND TEST SET. THE SAMPLING RATE OF THE ACCELERATION SIGNAL WAS 250 HZ

Method	Audio Sampling Rate		
	8,000 Hz	22,050 Hz	44,100 Hz
Audio only (RF)	0.6170	0.6885	0.7421
Audio only (CNN)	0.8630	0.8689	0.8981
Acc. only (RF)	0.5087		
Acc. only (CNN)	0.7997		
Multimodal (RF)	0.6517	0.6443	0.7953
Multimodal (CNN)	**0.9298**	**0.9347**	**0.9357**

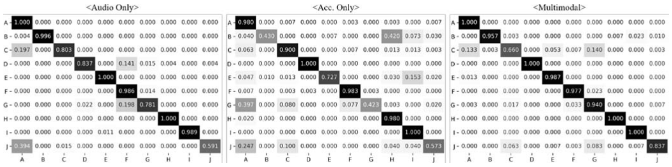
D. Deep Learning Approach

Among the various deep learning techniques, we adopted a CNN owing to its considerable time-series classification capability with a relatively low computational cost compared with a recurrent neural network (RNN) [10]. For audio feature extraction, a 64 × 64 melspectrogram image was generated from a raw signal and used as an input representation for the 2D-CNN model. To extract the acceleration features, a normalized signal was used as an input representation for the 1D-CNN model. The feature volumes from each modality were then concatenated followed by a fully connected (dense) layer. The detailed architecture of the network is shown in Fig. 4. To reduce the computational cost, we applied as simple a network structure as possible, ensuring a reasonable accuracy. The CNN was implemented in Python 3.6 using Keras (keras.io).

IV. EXPERIMENTS AND RESULTS

A. Accuracy for the Blind Test Set

To assess the performance of the proposed approach in a real-world situation, we collected additional data, namely, a blind test set. A total of 300 samples per class were collected for the same ten classes but for different objects with different ambient noises. We compared the test accuracy of the unimodal and multimodal approaches by varying the machine learning method and the sampling rate of the audio signal. The sampling rate of the audio signal was originally 44 100 Hz and then converted through a downsampling technique. Table II summarizes the results obtained. The training and testing were conducted using a personal computer (PC) with an NVIDIA GPU (Titan XP), running on the Linux (Ubuntu 18.04.3 LTS) operating system. Overall, it tended to achieve a higher test accuracy when using multimodality rather than unimodality, and when using a deep learning method rather than a shallow learning approach. The deep learning approach (CNN) demonstrated a better performance than the shallow learning method (RF). Despite being a sophisticated feature engineering process, the RF classifier was unable to achieve a high accuracy on the blind test set. Meanwhile, the test accuracy tended to increase as the sampling rate increased, but not significantly for the multimodal CNN approach. Overall, the proposed machine learning pipeline, i.e., multisensor fusion, outperformed the unimodality-based results. In particular, the test accuracy approached 93% even with the lower sampling rate of the audio signal (8000 Hz), indicating that the proposed method can be implemented on low-cost systems.



<Audio Only>

	A	B	C	D	E	F	G	H	I	J
A	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
B	0.004	0.996	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	0.197	0.000	0.803	0.000	0.000	0.000	0.000	0.000	0.000	0.000
D	0.000	0.000	0.000	0.837	0.000	0.141	0.015	0.004	0.000	0.004
E	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
F	0.000	0.000	0.000	0.000	0.000	0.986	0.014	0.000	0.000	0.000
G	0.000	0.000	0.000	0.022	0.000	0.198	0.781	0.000	0.000	0.000
H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
I	0.000	0.000	0.000	0.000	0.011	0.000	0.000	0.000	0.989	0.000
J	0.394	0.000	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.591

<Acc. Only>

	A	B	C	D	E	F	G	H	I	J
A	0.980	0.000	0.007	0.000	0.000	0.003	0.000	0.003	0.000	0.007
B	0.040	0.430	0.000	0.000	0.007	0.000	0.000	0.420	0.073	0.030
C	0.063	0.000	0.900	0.000	0.000	0.007	0.000	0.013	0.013	0.003
D	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
E	0.047	0.010	0.013	0.000	0.727	0.000	0.000	0.030	0.153	0.020
F	0.007	0.000	0.003	0.003	0.000	0.983	0.003	0.000	0.000	0.000
G	0.397	0.000	0.080	0.000	0.000	0.077	0.423	0.003	0.000	0.020
H	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.980	0.000	0.000
I	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
J	0.247	0.000	0.100	0.000	0.000	0.000	0.000	0.040	0.040	0.573

<Multimodal>

	A	B	C	D	E	F	G	H	I	J
A	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
B	0.000	0.957	0.003	0.000	0.000	0.000	0.000	0.007	0.023	0.010
C	0.133	0.003	0.660	0.000	0.053	0.007	0.140	0.000	0.000	0.003
D	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
E	0.013	0.000	0.000	0.000	0.987	0.000	0.000	0.000	0.000	0.000
F	0.000	0.000	0.000	0.000	0.000	0.977	0.023	0.000	0.000	0.000
G	0.003	0.000	0.017	0.000	0.000	0.033	0.940	0.000	0.000	0.007
H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
I	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
J	0.000	0.000	0.063	0.000	0.007	0.003	0.083	0.000	0.007	0.837

Fig. 5. Confusion matrices for the multimodal CNN approach using an audio sampling frequency of 44 100 Hz. Rows and columns denote the actual and predicted class labels, respectively.

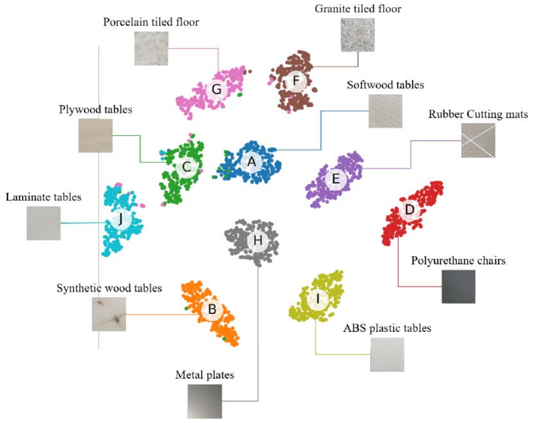


Fig. 6. t-SNE visualization of the last hidden layer representations in the multimodal CNN approach using the blind test set.

TABLE III
EXPERIMENTAL RESULTS: MEAN AND STANDARD DEVIATION OF THE MEASURED INFERENCE TIME

Machine	Audio Sampling Rate	Inference Time
Personal Computer	8000 Hz	36.83 ± 0.48 ms
	22050 Hz	38.53 ± 0.35 ms
	44100 Hz	41.96 ± 0.32 ms
Embedded Machine	8000 Hz	297.73 ± 3.64 ms
	22050 Hz	301.33 ± 2.81 ms
	44100 Hz	320.38 ± 2.49 ms

In Fig. 5, the confusion matrices when using a deep learning approach with an audio sampling at 44 100 Hz are shown. In terms of the audio system, classes C (plywood table), G (porcelain tiled floor), and J (laminate table) were difficult to discriminate, and classes B (synthetic wood table) and G (porcelain tiled floor) achieved a lower accuracy for the inertial system. The multisensor fusion architecture was able to achieve a high accuracy by complementing the misclassifications in each modality. We examined the internal features learned by the CNN using t-distributed stochastic neighbor embedding (t-SNE) [11], as shown in Fig. 6. Each point represents an input signal projected from the 128-D output of the last hidden layer of the CNN into two dimensions. We can see clusters of points of the same surface classes. In summary, all surface classes were well classified based on the proposed multimodal approach.

B. Implementation on Embedded Machine

To assess the practicality of the proposed method, we measured the inference time on an embedded environment. We duplicated the trained model (multimodal CNN) on an embedded machine, i.e., a Raspberry Pi 4 Model B. The measured time includes the following processes: the preprocessing of raw data, conversion into an input representation, feature extraction, and classification. Table III summarizes the mean and standard deviation of 100 repeated measures of the inference time. The inference time measured on a PC is also shown for comparison. The time increased as the audio sampling rate increased. The fastest time was approximately 298 ms with an audio sampling rate of 8000 Hz. Because the proposed system does not require a consecutive inference in a real-time manner, it can be successfully applied to practical embedded applications, such as AI speakers.

V. CONCLUSION

In this letter, we proposed a multivariate time-series classification system that fuses heterogeneous sensor measurements using a late fusion CNN. By listening to the multivariate measurements, i.e., sound and inertial signals, when hitting a surface, the proposed system can identify ten different categories of surfaces found in our daily environment. For the blind test set, a test accuracy of approximately 93% was achieved even with a lower sampling rate (i.e., 8000 Hz). In addition, because the inference time was measured at less than 300 ms, we believe that our system can be applied to various embedded devices, enabling a contextual human–machine interaction. Future studies will focus on extending the proposed approach to more diverse surfaces and implementing in real-world applications with the further miniaturization of the system.

REFERENCES

[1] Z. Zhou, Y. Li, J. Liu, and G. Li, "Equality constrained robust measurement fusion for adaptive kalman-filter-based heterogeneous multisensor navigation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 4, pp. 2146–2157, Oct. 2013.

[2] S. Safari, F. Shabani, and D. Simon, "Multirate multisensor data fusion for linear systems using kalman filters and a neural network," *Aerosp. Sci. Technol.*, vol. 39, pp. 465–471, Dec. 2014.

[3] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio–visual temporal binding for egocentric action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 5492–5501.

[4] L. Shi, M. Ashoori, Y. Zhang, and S. Azenkot, "Knock knock, what's there: Converting passive objects into customizable smart controllers," in *Proc. 20th Int. Conf. Human Comput. Interact. Mobile Devices Serv.*, 2018, p. 31.

[5] J. Cho, I. Hwang, and S. Oh, "Vibration-based surface recognition for smartphones," in *Proc. IEEE Int. Conf. Embedded Real Time Comput. Syst. Appl.*, Seoul, South Korea, 2012, pp. 459–464.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[7] T. Gong, H. Cho, B. Lee, and S.-J. Lee, "Knocker: Vibroacoustic-based object recognition with smartphones," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 3, p. 82, 2019.

[8] V. Radu *et al.*, "Multimodal deep learning for activity and context recognition," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–27, 2018.

[9] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[11] L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.