

Shape-N-Motion: Fine-Grained Hand Object Manipulation Recognition with Ultrasonic and IMU

Kaito Fujishige*, Kota Tsubouchi†, Yuuki Nishiyama‡, Masamichi Shimosaka*

*Institute of Science Tokyo, Tokyo, Japan

Email: {fujishige, simosaka}@miubiq.cs.titech.ac.jp

†LY Corporation, Tokyo, Japan

Email: ktsubouc@lycorp.co.jp

‡The University of Tokyo, Tokyo, Japan

Email: nishiyama@sis.u-tokyo.ac.jp

Abstract—Fine-grained hand-object manipulations (e.g., tapping vs. scrolling on the same smartphone) are informative for lifelogging and remote health monitoring, but are often confused by wearable IMU (Inertial Measurement Unit) and audible audio because wrist trajectories are similar and many actions are nearly silent. We propose *Shape-N-Motion*, a wrist-worn multimodal system that fuses fingertip-oriented ultrasonic echoes and a wrist IMU for direct time-series classification without hand-shape reconstruction. We build a 45-class benchmark (N=6) refined from existed protocol to explicitly reflect contact states, state changes, and tool use, and evaluate cross-participant generalization with LOPO (Leave-One-Participant-Out). *Shape-N-Motion* achieves 0.467 accuracy / 0.417 Macro-F1, outperforming the strongest IMU+audio baseline by +0.081 accuracy and +0.079 Macro-F1.

Index Terms—Wearable sensing, ultrasonic sensing, IMU, activity recognition, spectrogram features.

I. INTRODUCTION

Fine-grained hand-object manipulations reflect subtle differences in fingertip motion and contact state and are therefore promising signals for applications such as remote health monitoring and daily assistance for older adults [1]. However, recognizing actions performed on the *same* object (e.g., browsing vs. typing on a smartphone) remains challenging. Prior wearable systems based on IMU and audible audio frequently confuse these actions because (i) IMU captures mainly coarse wrist/arm trajectories and cannot directly observe fingertip contact, and (ii) audible audio is unreliable for silent actions and everyday background noise [2], [3].

Ultrasonic sensing provides an opportunity to capture millimeter-scale hand-shape and contact cues through high-frequency reflections [4], [5]. Nevertheless, many ultrasonic pipelines first reconstruct hand pose; under object contact, multipath scattering distorts echoes and degrades reconstruction accuracy, which can destabilize downstream recognition [5]. We therefore introduce *Shape-N-Motion*, a multimodal sensor system that fuses a fingertip-oriented ultrasonic sensor with a wrist-mounted IMU sensor. We advocate a *direct classification* strategy that directly analyzes multimodal time-series captured by ultrasonic echoes and a wrist IMU. By avoiding any intermediate reconstruction stage, it learns manipulation-specific local reflection patterns for recognition.

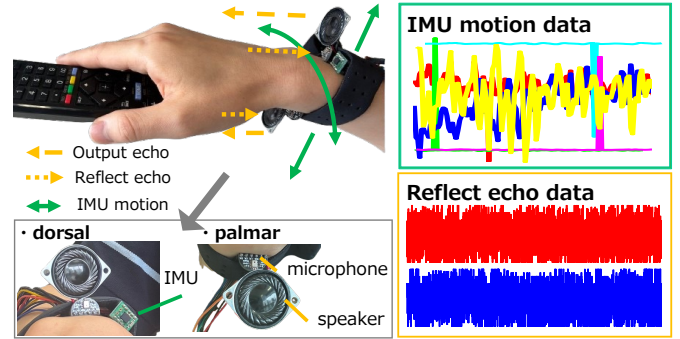


Fig. 1. Developed wrist module: dorsal/palmar ultrasonic transducers and a dorsal IMU, driven by an onboard microcontroller

This research makes the following contributions:

- We formulate fine-grained hand-object manipulation recognition focusing on confusing micro-manipulations on the same object.
- We design a wrist-worn sensing layout that combines fingertip-oriented ultrasonic echoes with a wrist IMU and performs recognition without hand-shape reconstruction.
- We demonstrate improved discrimination on an 45-class benchmark, including reduced confusion between similar smartphone actions compared with IMU+audio baselines.

II. RELATED WORK

A. Wearable Object Manipulation Recognition

Wearable recognition has explored combinations of IMU, audio, and vision to identify daily activities and object manipulations [2], [6]. While coarse activities can be recognized reliably, distinguishing actions on the same object remains difficult. Bhattacharya *et al.* reported about 20% mutual confusion between smartphone browsing and typing using IMU+audio [2], highlighting the need for sensing modalities that capture fingertip-level interaction cues beyond wrist motion.

B. Ultrasonic Sensing for Hand Shape and Interaction

Ultrasonic sensing has been used for finger tracking, hand pose estimation, and wrist-worn interaction sensing [4], [5], [7]. These studies demonstrate that high-frequency echoes

can encode fine hand-shape information. However, when the hand contacts an object, multipath reflections increase echo complexity and can degrade pose reconstruction [5]. This motivates recognition approaches that *exploit* contact-induced echo changes rather than treating them as noise.

III. PROBLEM FORMULATION AND MOTIVATION

A. Formulation

We address manipulation recognition as a multi-class classification problem. Given synchronized multimodal time-series from sensors $\mathcal{S} = \{\text{Ultrasonic}, \text{IMU}\}$ over a window of length T , we predict a manipulation label $y \in \{1, \dots, C\}$. For modality $s \in \mathcal{S}$, the observation sequence is $\mathbf{X}^{(s)} \in \mathbb{R}^{T \times d_s}$. A modality-specific encoder $\phi^{(s)}(\cdot)$ produces an embedding $\mathbf{h}^{(s)} = \phi^{(s)}(\mathbf{X}^{(s)})$. A fusion module $g(\cdot)$ yields a fused representation $\mathbf{z} = g(\mathbf{h}^{(\text{Ultrasonic})}, \mathbf{h}^{(\text{IMU})})$, and a classifier predicts $\hat{y} = f(\mathbf{z})$. All parameters are trained end-to-end with cross-entropy loss.

B. Why Ultrasonic Echoes Help

IMU captures arm/wrist kinematics at high temporal resolution but cannot directly observe fingertip contact states, making it inherently insufficient for manipulations with similar trajectories [2], [8]. Audible audio often provides weak cues for silent manipulations and is sensitive to ambient noise [3]. In contrast, ultrasonic echoes around 19–22 kHz are influenced by local geometry and contact; they exhibit manipulation- and object-dependent spectral patterns (micro-Doppler and attenuation) that remain informative even when actions are nearly silent [5].

IV. PROPOSED METHOD: SHAPE-N-MOTION

A. Sensor Layout and Implementation

A wrist-worn prototype mounts (i) a pair of ultrasonic channels (speaker + microphone) on opposite sides of the wrist, oriented toward the fingertips, and (ii) an IMU on the dorsal wrist (see Fig. 1). Ultrasonic transducers acquire local reflection patterns shaped by fingertip pose/contact, while the IMU provides complementary motion cues for actions involving similar hand shapes but different trajectories.

B. Signal Processing

Ultrasonic: We transmit a 19.0–22.0 kHz correlation-based FMCW chirp with a 0.012 s repetition period and record echoes. We compute the STFT with a 0.012 s window/hop and extract 64-D power-spectrum features per channel over 19.0–22.0 kHz, corresponding to the transmitted ultrasonic band. We concatenate the L/R channels into a 128-D vector per frame.

IMU: 3-axis acceleration and 3-axis angular velocity are sampled at 50 Hz and aligned to the STFT frame timestamps via on-board clocks. Missing values are imputed with zero-order hold per STFT frame (0.012 s).

For both modalities, we extract a 3 s window centered at the midpoint of each annotated segment. If the window extends beyond the signal boundaries, we pad the missing part with the boundary values (replication padding).

C. Model and Fusion

Figure 2 illustrates the architecture. The IMU branch uses depthwise–pointwise convolutions followed by a Transformer encoder to capture long-range motion patterns. The ultrasonic branch uses 1D convolution, BiLSTM, and attention pooling to emphasize local reflection changes under contact. Branch outputs are weighted by a temperature-scaled softmax ($\tau = 2.0$), concatenated, and classified via a fully connected layer. We optimize with Adam ($\text{lr } 1 \times 10^{-3}$, batch 64) using cross-entropy loss.

V. EXPERIMENT

A. Hardware

Both the ultrasonic module and the IMU are controlled by a Raspberry Pi 4 Model B. The ultrasonic channel consists of an ultrasonic speaker (CMS-28528N-L152) and a MEMS microphone (MSM261S4030H0), driven by a Class-D amplifier (MAX98357A). The IMU is a Bosch BNO055; we use six inertial channels (3-axis acceleration and 3-axis angular velocity).

B. Benchmark design and evaluation protocol

Participants ($N=6$) wear the device on the dominant wrist and perform hand–object manipulations in everyday environments. Following the 23-class protocol by Bhattacharya *et al.*, we annotate temporal segments using synchronized RGB videos and refine the labels into a 45-class taxonomy (Fig. 3) that explicitly reflects **contact states, state changes, and tool use**. This refinement yields challenging cases where motions are similar but interaction states differ, requiring fine-grained interaction cues beyond wrist dynamics.

Each action consists of “grasp an object, perform the designated manipulation, and return to the start position.” One session is a 2-minute continuous recording; the target action is repeated roughly once every 20–30 seconds, and we collect three sessions per action per participant.

We evaluate cross-participant generalization using LOPO (leave-one-participant-out) and report the mean over five folds.

C. Baselines and metrics

We compare against two strong IMU+audio baselines [2] (Late fusion + Random Forest; Late fusion + Attend & Discriminate (CNN14)) and an ultrasonic-only baseline (EchoWrist [5]). For fair comparison, the IMU+audio baselines are down-sampled to remove the ultrasonic band. We report Accuracy and Macro-F1, and inspect confusion trends between similar classes.

D. Results

Table I summarizes results on the 45-class fine-grained benchmark ($N=6$, LOPO). Shape-N-Motion outperforms all baselines in both Accuracy and Macro-F1. These gains are most pronounced for classes with similar wrist motion but different contact or state changes, where ultrasonic echoes provide additional discriminative cues.

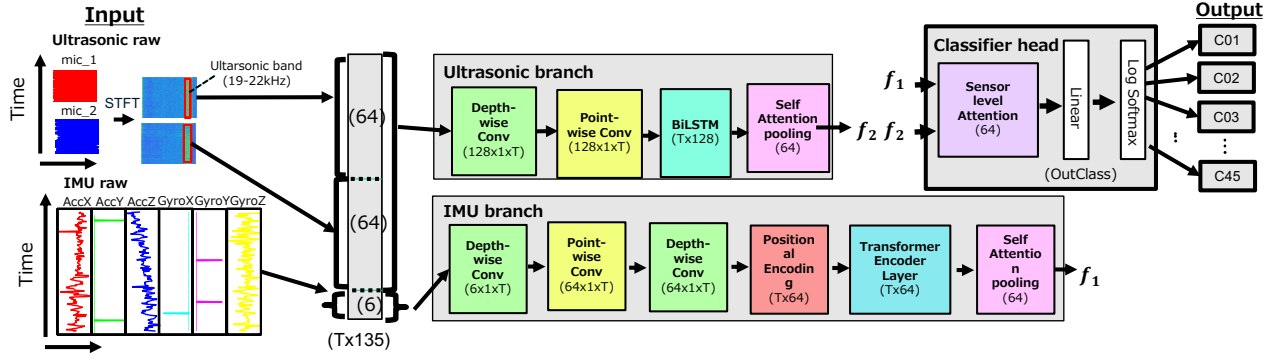


Fig. 2. Model overview: modality-specific encoders and temperature-scaled fusion.



Fig. 3. Our 45-class fine-grained hand-object manipulation benchmark.

TABLE I
RECOGNITION PERFORMANCE ON THE 45-CLASS BENCHMARK (N=6, LOPO).

Method	Acc.	Macro-F1
Shape-N-Motion (Proposed)	0.467	0.417
Random Forest	0.228	0.143
Attend & Discriminate	0.386	0.338
EchoWrist	0.157	0.124

VI. DISCUSSION

A. Why ultrasonic cues help in the 45-class setting

The 45-class setting contains many pairs with similar wrist trajectories but different interaction states. While IMU + audio mainly capture motion dynamics, ultrasonic echoes capture interaction-dependent local reflection changes, reducing confusion among similar motions and improving Macro-F1. Thus Ultrasonic cues help distinct such complicate similar pairs.

B. Limitations

We do not exhaustively cover wearing-position variations, object geometries, or environmental changes; broader validation and personalization remain future work.

VII. CONCLUSION

We presented Shape-N-Motion, a wrist-worn multimodal system that fuses ultrasonic echoes and IMU signals to recognize fine-grained hand-object manipulations without hand-shape reconstruction. On the 45-class fine-grained benchmark (N=6, LOPO), the proposed method achieves Accuracy 0.467 / Macro-F1 0.417, outperforming strong IMU+audio baselines and EchoWrist. These results suggest that ultrasonic sensing

provides cues about contact and state changes through local reflections, strengthening discrimination among similar motions.

REFERENCES

- [1] J. Zhang, Y.-J. Zhao, J.-Y. Wang, H. Cui, S. Li, X. Meng, R.-Y. Cai, J. Xie, S.-Y. Sun, Y. Yao *et al.*, “Comprehensive assessment of fine motor movement and cognitive function among older adults in china: a cross-sectional study,” *BMC geriatrics*, vol. 24, no. 1, p. 118, 2024.
- [2] S. Bhattacharya, R. Adaimi, and E. Thomaz, “Leveraging sound and wrist motion to detect activities of daily living with commodity smartwatches,” *Proc. of ACM IMWUT*, vol. 6, no. 2, pp. 1–28, 2022.
- [3] D. Liang and E. Thomaz, “Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos,” *Proc. of ACM IMWUT*, vol. 3, no. 1, pp. 1–18, 2019.
- [4] Q. Zhang, Y. Lin, Y. Lin, and S. Rusinkiewicz, “Hand pose estimation with mems-ultrasonic sensors,” in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–11.
- [5] C.-J. Lee, R. Zhang, D. Agarwal, T. C. Yu, V. Gunda, O. Lopez, J. Kim, S. Yin, B. Dong, K. Li, M. Sakashita, F. Guimbretiere, and C. Zhang, “Echowrist: Continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband,” in *Proc. of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–21.
- [6] E. J. Wang, T.-J. Lee, A. Mariakakis, M. Goel, S. Gupta, and S. N. Patel, “Magnifisense: Inferring device interaction using wrist-worn passive magneto-inductive sensors,” in *Proc. of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 15–26.
- [7] Y. Shi, C. Yu, X. Lu, X.-D. Yang, Y. Wang, and Y. Shi, “Sonarwatch: Field sensing technique for smartwatches based on ultrasound and motion,” *arXiv preprint arXiv:2408.12689*, 2024.
- [8] J. Ranjan and K. Whitehouse, “Towards recognizing person-object interactions using a single wrist wearable device,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp. 722–731.