ViObject: Harness Passive Vibrations for Daily Object Recognition with Commodity Smartwatches

WENQIANG CHEN, Chinese Academy of Sciences, SIAT; Massachusetts Institute of Technology, USA
SHUPEI LIN, VibInt AI, China
ZHENCAN PENG, VibInt AI, China
FARSHID SALEMI PARIZI, University of Washington, USA
SEONGKOOK HEO, University of Virginia, USA
SHWETAK PATEL, University of Washington, USA
WOJCIECH MATUSIK, Massachusetts Institute of Technology, USA
WEI ZHAO, University of SIAT, Shenzhen, China
JOHN STANKOVIC, University of Virginia, USA

Knowing the object grabbed by a hand can offer essential contextual information for interaction between the human and the physical world. This paper presents a novel system, ViObject, for passive object recognition that uses accelerometer and gyroscope sensor data from commodity smartwatches to identify untagged everyday objects. The system relies on the vibrations caused by grabbing objects and does not require additional hardware or human effort. ViObject's ability to recognize objects passively can have important implications for a wide range of applications, from smart home automation to healthcare and assistive technologies. In this paper, we present the design and implementation of ViObject, to address challenges such as motion interference, different object-touching positions, different grasp speeds/pressure, and model customization to new users and new objects. We evaluate the system's performance using a dataset of 20 objects from 20 participants and show that ViObject achieves an average accuracy of 86.4%. We also customize models for new users and new objects, achieving an average accuracy of 90.1%. Overall, ViObject demonstrates a novel technology concept of passive object recognition using commodity smartwatches and opens up new avenues for research and innovation in this area.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; Object Recognition; Accessibility technologies.

Additional Key Words and Phrases: Wearable Sensing, Object Recognition, Tangible Interaction, Vibration Sensing

ACM Reference Format:
Wenqiang Chen, Shupei Lin, Zhencan Peng, Farshid Salemi Parizi, Seongkook Heo, Shwetak Patel, Wojciech Matusik, Wei Zhao, and John Stankovic. 2024. ViObject: Harness Passive Vibrations for Daily Object Recognition with Commodity Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 5 (2024), 26 pages. https://doi.org/10.1145/3643547

Authors' addresses: Wenqiang Chen Chinese Academy of Sciences, SIAT; Massachusetts Institute of Technology, USA, wenqiang@mit.edu; Shupei Lin VibInt AI, China; Zhencan Peng VibInt AI, China; Farshid Salemi Parizi University of Washington, USA; Seongkook Heo University of Virginia, USA; Shwetak Patel University of Washington, USA; Wojciech Matusik Massachusetts Institute of Technology, USA; Wei Zhao University of SIAT, Shenzhen, China; John Stankovic University of Virginia, USA.

2474-9567/2024/3-ART5 $15.00
https://doi.org/10.1145/3643547

Fig. 1. ViObject is a system that recognizes untagged everyday objects using accelerometer and gyroscope sensor data from commodity smartwatches. Passive object recognition is accomplished by identifying the vibrations caused by grab-objects without the need for additional hardware or human effort.

1 INTRODUCTION

Despite the massive advancements in technology today, there still remains a distinct divide between our physical world and the technological world. While most of us interact with electronic devices on a daily basis, we exist in two separate realms in a sort of symbiosis, where we benefit from the convenience of technology while technology builds and improves with our use over time. However, it is still easy to point at an object and identify which realm it belongs to, the technological world or our own. In this work, we explore a world heavily inspired by the idea of changing the world itself into an interface [34].

Daily object recognition is one of the technologies to help bridge the divide between our world and the technological world by using mundane objects as triggers for specific services. Attaching tags to objects has been widely proposed, where tags are used to retrieve information about the objects. QR codes, RFID tags [55], near-field communication (NFC) [22], and acoustic barcodes [29] have been utilized to recognize and automatically select the target service from the mobile devices. Vision-based solutions utilize computer vision and machine learning techniques to identify objects captured within the frame of the cameras [19, 35]. Capacitivo [58] and Tessutivo [23] recognized objects placed on customized fabrics, using capacitive sensing. Electromagnetic (EM) based sensing approaches require specialized EM sensors as well, and are applicable to only electrical appliances that emit EM signals [40, 53]. Other approaches also require customized devices to emit active signals (e.g., vibrator [46], millimeter wave[64], etc) to identify objects. Despite these advantages, customized devices reduce accessibility and thus their deployability is limited. Some acoustic-based approaches recognize objects through different sounds by knocking on different objects [24, 43, 51]. However, these approaches require users to have extra actions to indicate users' intentions, which disrupts people's daily activities and sets a boundary between the technology world and the physical world.

This paper introduces ViObject, a system designed for tangible interactions that harnesses passive vibrations—originating from the grasp of everyday objects—instead of active vibrations produced by vibrators, to

enable object recognition. By capturing the unique passive vibrations that propagate through the hand when an object is grabbed, ViObject can identify untagged everyday objects. Leveraging the accelerometer and gyroscope sensors embedded in commodity smartwatches, ViObject can capture and process these grab-object-induced vibrations, enabling seamless and intuitive object recognition without requiring additional hardware or user effort.

One of the key advantages of ViObject is its focus on smooth integration with people's daily lives. Unlike other techniques that require users to take extra actions, such as attaching sensors or taking pictures, ViObject reads data from users' daily movements using a hand-worn smartwatch without disrupting their habits or daily routines. This facilitates borderless and fluid interactions between the technological world and our daily lives. Additionally, the popularity of commodity smartwatches and the development community surrounding them makes ViObject a practical and accessible solution for a wide range of applications.

The development of ViObject poses several challenges that need to be addressed. Firstly, the vibration signal generated by grabbing an object is relatively weak and can be overwhelmed by the hand's movement signal, particularly with a limited Inertial Measurement Unit (IMU) sampling rate. Moreover, unlike active sensing approaches that use customized signals with specific frequencies as a signal source for sensing, the passive signals captured from grabbing objects are unmodulated and have varying frequencies, which makes detection and feature extraction more challenging. Additionally, the vibration signal induced when the user touches an object can vary depending on applied pressure and speed. Objects vary in shape and size, so the induced vibration signals also depend on object-touching positions. Lastly, new end users may need to identify new objects that are unique to their homes.

To overcome these challenges, we propose several techniques in ViObject. Firstly, we eliminate interference through signal processing and augmentation techniques. Secondly, we use interpolation to enhance samples and leverage attention-based residual networks to improve the system's accuracy. Thirdly, we design an adversarial training regularization with center loss to mitigate the impact of orientation changes. Finally, we employ generalized few-shot learning with data synthesis for object customization, enabling the system to recognize new objects and users with minimal training data. These techniques enable ViObject to effectively recognize objects passively using smartwatches' accelerometer and gyroscope sensor data, addressing new research and innovation opportunities.

The performance of ViObject was evaluated using data collected from 20 participants interacting with 20 different daily objects while wearing a smartwatch. The collected data was analyzed to assess the system's basic performance, and the average recognition accuracy was found to be 86.4%. A subsequent experiment was conducted with 10 additional participants to test ViObject's ability to recognize new objects customized by users. The results demonstrated excellent accuracy (90.1%) for new users and new objects, even when the objects were grabbed from different angles with varying pressure and speed, and different parts of the object were grabbed. The performance of ViObject was consistent across different smartwatches and over a week. An end-to-end standalone app was implemented on an Android smartwatch for object recognition in real-time, which included the Android TextToSpeech module to play the sound of the prediction. A user experience study was conducted to evaluate the app, and the results showed positive user feedback. Furthermore, ViObject achieved low latency and low power consumption (see section 5), making it a practical and efficient solution for everyday use.

This system has its inherent limitations. Numerous objects have yet to be tested in real-world scenarios, especially when objects bear close resemblance to each other. As the variety of objects increases, distinguishing between some may pose challenges. Nevertheless, ViObject strives to validate a new concept: recognizing objects from daily grasp actions. Moreover, most applications only need to employ a selective and manageable set of objects. For instance, recognizing a single pill bottle can facilitate medication reminders, while identifying a singular dumbbell can support fitness tracking. A curated set of daily objects could be instrumental in assessing Alzheimer's disease progression. A few specified daily objects can also activate corresponding smart home

services. In escape room scenarios, ViObject can detect when participants pick up a few particular objects, subsequently triggering clues or furnishing additional insights via the smartwatch to aid puzzle-solving. By identifying passive vibrations from daily object interactions, ViObject bridges the gap between our physical and digital realities, transforming the physical world around us into an interactive interface.

To summarize, our main contributions are:

- To the best of our knowledge, we are the first to harness passive vibrations to recognize daily objects using the IMU sensor found in many common smartwatches. We also conducted a series of feasibility studies to understand the principle behind this new concept.
- We have designed a novel technical process to eliminate interference (e.g., hand movement, grasp speed, and strength), extract fine-grained features from various object-touching positions using a transformer-based network, and customize the model for new users and new objects.
- We have developed an end-to-end standalone system in commercial smartwatches, which can achieve real-time object recognition. Extensive experiments have been conducted to demonstrate ViObject's accuracy, robustness, and user experience.

2 RELATED WORK

We first review prior sensing systems and methods that enable similar object recognition capabilities as ViObject. We then cover work that more closely aligns with our technical approach of vibration sensing.

2.1 Vibration Based Object Recognition

A mechanical vibration that propagates through a medium leaves a unique signature. Vibration-based object recognition follows two strategies: active vibration and passive vibration.

The former is to have a device generating modulated active vibrations, e.g., motor-powered vibrations. For example, ViBand [39] measured a vibration transmitted through the human body to recognize the object which attached a tag generating vibrations. VibSense [42] attached a vibrator on objects. VibEye [46] made devices worn on fingers to generate vibrations. In this way, this system can recognize objects when users hold objects. Unlike these existing works which only work with devices generating active vibrations, ViObject does not require devices to make vibrations but uses passive vibrations induced by grabbing objects.

The other approaches [24, 43, 51] recognized objects through passive sounds by knocking on different objects. However, these approaches require users to have extra actions to indicate users' intentions, which disrupts people's daily activities and sets a boundary between the technology world and the physical world. For instance, Knocker [24] introduced a method enabling users to tap their phone on a target for object identification, akin to using a phone to capture an image for object recognition. However, both these approaches necessitate additional user actions, potentially disrupting their ongoing activities. In contrast, ViObject identifies user's hand grasp without requiring extra actions to their daily activities. Moreover, while Knocker assumes users must tap on a specific spot of an object, ViObject explores diverse object grasp variations, particularly different positions of contact. To solve these challenges, ViObject first mitigates interference through adept signal processing and augmentation techniques. Secondly, it employs interpolation to enrich sample data and leverages attention-based residual networks to bolster the system's accuracy. Thirdly, an adversarial training regularization combined with center loss design is implemented to alleviate the impact of orientation changes.

ViObject aims to help remove these currently necessary extra actions and aims to create borderless and fluid interactions between the technological world and our own. Thus, ViObject recognizes grab-object-induced vibrations without affecting our daily activities.

2.2 Other Object Recognition Methods

There are many technologies using different sensors for sensing, such as cameras [19, 35], capacitive sensors [48], RF sensors [20, 64], light sensors [27, 49, 65], EMG sensors [21]. Ohnishi et al. [47] captured images using a wrist-worn camera to identify handheld objects. WristSense [44] recognized hand-held objects using wrist-worn cameras. Magic Finger [63] embedded multiple image sensors to track and capture the surfaces of different materials touched by a finger. However, vision-based systems are easily affected by the lighting condition of the environment and the misalignment of the object in the line of sight or require special hardwares such as RGB-Depth cameras. RadarCat [64] is a technique using millimeter-wave radar. It could classify various kinds of objects such as body parts, transparent materials, and everyday objects placed on the sensing pad. TUIC [66] identified tagged objects through a touchscreen (i.e. mutual capacitive sensing) by recognizing the geometrical pattern of the tag. With Capacitive NFCs [25], target objects were instrumented using an active tag. Zanzibar [52] recognized objects by implementing NFC tags. TagScan [54] identified different types of liquids using RFID. Another approach in recognizing and controlling appliances is sensing electromagnetic (EM) emissions [40, 59]. This approach exploits the uniqueness of EM signals emitted by electronic appliances for object recognition. The EM-based approach is, however, limited to electronic appliances since non-electronic objects do not emit EM signals. Most recently, Capacitivo [58] and Tessutivo [23] recognized objects placed on fabrics, using capacitive sensing. These works either attached tags on objects or used customized hardware. Unlike these works, ViObject recognizes untagged daily objects by passive grab-induced vibrations with a commodity smartwatch, which is more practical and easily accessed than customized devices.

2.3 Vibration Based Sensing

Vibrations carry rich information of locations [8–10, 12, 17], identifications [4, 26, 57, 57, 62], gestures [5–7, 14–16, 28, 67], activities [11, 13, 33], speech [26], materials [3, 46], weights [69] and so on for sensing technology. Kunze et al. [37] and Cho et al. [18] find the location of a mobile phone by measuring acceleration (and also sound [37]) using internal sensors after imposing a vibration. Such hand-transmitted vibrations are generated and measured by surface transducers and contact microphones. Most recently, ViFin [5] harnesses passive finger movement vibration to recognize in-air micro finger writing. There are some works [4, 8, 68] recognizing locations of the human body where a finger taps on it. These systems learned the unique profile of the finger-tap-induced vibrations on a specific body area. TouchPass [61] authenticated users by detecting touch vibrations on smartphones. VibeBin [71] took advantage of resonance when the object was exposed to vibration. This system learned the discrete fill-levels of a waste bin and then classified them using clustering. Some researchers used the motor in smartphones to generate active vibrations. VibroScale [69] estimates the relative induced intensity of an object placed on a smartphone. Vi-Liquid [32] identified different liquids in a container placed on a smartphone. Vibrosight [70] remotely detected unique vibration patterns of activities with laser vibrometry.

Compared to these works, our work is the first to harness grab-object-induced passive vibrations to recognize different objects in daily life with a commodity smartwatch. ViBand [39] and the following works [38, 60] recognized various gestures using smartwatches. However, our system involves the interaction between hands and objects, which is more challenging than just hands. Sharma, etc. [50] classified hand gestures while holding an object. However, our work differs in approach and research objectives. Additionally, object recognition has different challenges from gesture recognition, such as different object-touching positions and different interference.

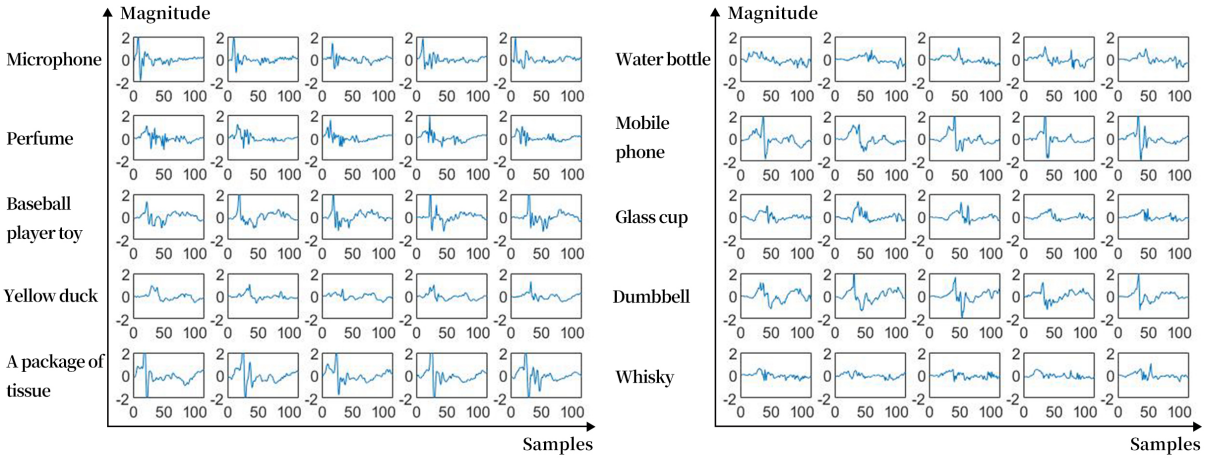

Fig. 2. Example signals of grab-induced vibrations from 10 daily objects (five examples of each) (max of accelerometer axes shown). The grab-object-induced vibration features from the same object are similar, while different from different objects. Photos of these objects are shown in Figure 10.

3 FEASIBILITY STUDY

In this study, we begin by explaining the theory of operations for our system. We then examine the uniqueness of objects by analyzing their vibration signals. Next, we investigate the factors that contribute to an object's uniqueness, including material differences, mass, size, and shape. Finally, we explore potential sources of interference for object uniqueness, such as variations in object-touching position, grasp speed, and strength.

3.1 Theory of Operations

The proposed system for passive object recognition is based on the physical phenomenon of mechanical vibration. When an untagged everyday object is grasped, the action of the fingers and hand generates vibration waves that propagate through the object and the hand, and are eventually detected by the IMU sensor in the smartwatch. Mathematically, the vibration signal can be modeled as a mechanical wave with a certain frequency, amplitude, and phase, which depends on the properties of the object, such as its mass, materials, and shape. The vibration signal can be expressed as a function of time, position, and direction using the wave equation:

$$\nabla^2 \vec{u} - \frac{1}{c^2} \frac{\partial^2 \vec{u}}{\partial t^2} = \vec{0} \tag{1}$$

where $\vec{u}$ is the displacement vector, c is the wave speed, and ∇^2 is the Laplace operator. The solution of this equation gives the vibration waveform that can be measured by the IMU sensor. The physics behind this approach is that the vibration signature of each object is unique and can be used as a fingerprint for identification. [24, 46] Overall, the system leverages the physical properties of mechanical vibration and mathematical tools to achieve passive object recognition using commodity smartwatches.

3.2 The Vibration Uniqueness of Objects

We conducted an experiment to observe the passive vibrations induced by grabbing objects. An experimenter grabbed 10 different daily objects (five times each) while wearing a commodity smartwatch. We plotted the vibration signals of the max of the accelerometer axes. As shown in Fig. 2, the grab-object-induced vibration features from the same object show similar patterns, while different from different objects. Photos of these

objects are shown in Figure 10. **Therefore, this experiment demonstrated that passive grab-object-induced vibrations are unique for different daily objects.**

3.3 Mass, Material, Size, or Shape?

Although the passive vibrations induced by different daily objects are unique, we investigated what enables ViObject to distinguish between different objects. Is it the material, weight, shape, or size of the object?

To answer this question, we conducted an experiment with four types of objects, as shown in Figure 4. Each object type consisted of four objects with different weights (Figure 4(1)): four plastic cubes containing different numbers of small hollow plastic cubes; materials (Figure 4(2)): aluminum, copper, wood, and iron; shapes (Figure 4(3)): four toy characters made of rubber with different shapes; and sizes (Figure 4(4)): air balls with different sizes. For each group of four objects, one factor was varied while the other three factors were kept constant. For example, air balls of different sizes had similar weights, shapes, and materials because the air inside is very light. We randomly grabbed each object 40 times.

To understand how vibration responses distinguish between objects, we used t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the vibration responses of different objects. t-SNE is a dimensionality reduction technique that visualizes data from high-dimensional to low-dimensional space in a way that similar points are plotted nearby. Figure 3 shows t-SNE for the four types of objects, with 40 samples per object (iteration=1K, perplexity=30).

Figure 3(1) shows that four objects with different weights are very close to each other, making it difficult to classify them. However, Figure 3(2) shows that four objects with different materials are easier to separate from each other, even though these four objects have the same shape and size. Therefore, we believe that material contributes significantly to the features of vibrations induced by grabbing objects. Figures 3(3) and 3(4) show that four objects with different shapes or sizes are well-separated with the most distinguishable distances, even though they have the same materials. We believe that different sizes and shapes of objects cause different finger gestures when grabbing them, which leads to different vibration profiles. **In conclusion, this experiment demonstrated that different masses could not differentiate between objects, while shape, size, and material jointly contribute to the recognition of objects through vibrations induced by grabbing.**

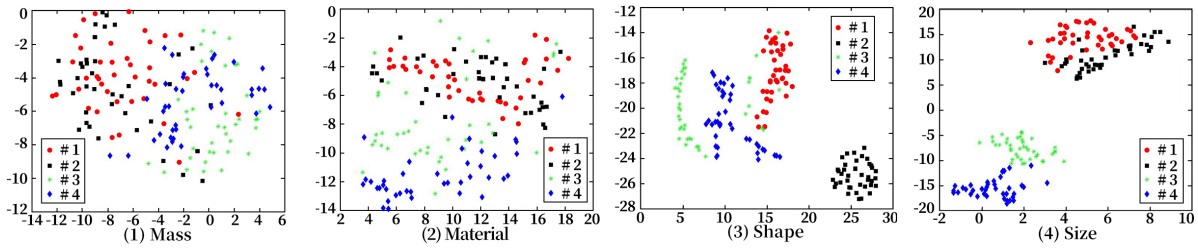

Fig. 3. T-SNE visualization of vibration responses for different objects. (1) Four objects with different weights. (2) different materials. (3) different shapes. (4) different sizes.

3.4 The Impact of Object-touching Position

In order to investigate the potential impact of touching different parts of an object on the vibration responses, we conducted a study using Pikachu (as shown in Fig. 4 (3)). Specifically, we conducted a total of 90 grabs of Pikachu, with each position being grabbed 30 times while maintaining a fixed object position. The positions for grabbing were changed three times during the study.

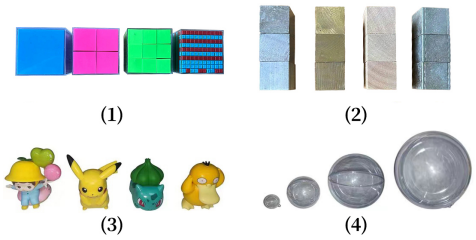

Fig. 4. Different (1) materials, (2) weights, (3) shapes, (4) sizes. These four sets of objects were used to study what reasons contribute to object recognition using grab-induced vibrations.

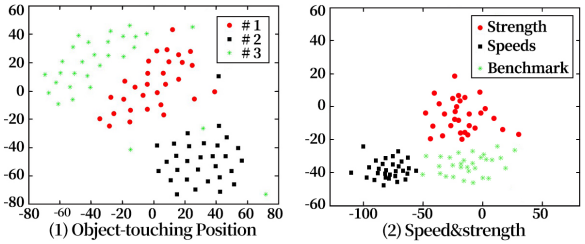

Fig. 5. T-SNE Visualization for (1) different object-touching positions and (2) different grab speeds and forces, highlighting the potential impact of these factors on the performance of object recognition.

We employed t-SNE to visualize the potential variations in vibration responses resulting from different object-touching positions, The results, as shown in Fig. 5 (1), confirmed our initial concern that different object-touching positions can indeed lead to variations in vibration responses. This finding highlights the need to address this potential problem in the object recognition system under development. Further studies have been conducted in the section 4.3 to develop appropriate methods (adversarial training regularization with center loss) to mitigate the variations resulting from different object-touching positions.

3.5 The Speed & Strength Interference

Building upon the aforementioned experiment, we designed a similar study to investigate the potential impact of the speed and force of object grabbing on the performance of the object recognition system.

To conduct the study, we used the same Pikachu object as before and varied the speed and force of the grabs while maintaining a fixed object position. Specifically, we conducted 60 grabs in total, with 30 grabs at higher speed and 30 grabs with stronger pressure.

We then employed t-SNE to visualize the potential variations in vibration responses resulting from different grab speeds and forces. The results, as shown in Fig. 5 (2), revealed that variations in grab speed and force can indeed lead to significant differences in vibration responses. This finding underscores the importance of accounting for the potential impact of grab speed and force on the performance of the object recognition system. We have developed data augmentation methods in section 4.1.2 to improve the system's performance under varying grab speed and force conditions.

3.6 The Impact of Different Users

In order to delve deeper into the variations, we designed an additional experiment by grasping five common items in Fig. 2 with more variations. Initially, we obtained baseline signals from these objects, establishing a reference for subsequent comparisons. To unravel the nuances within these vibrations, our experiment deliberately introduced variations in grasp speed (fast), position, strength (gentle), and involved different users across the objects (refer to Fig. 6). This intentional manipulation simulated the diverse ways individuals naturally interact with items in their daily lives. These nuanced interactions yielded distinct vibration signals, forming a comprehensive dataset that underscored intricate differences among objects and highlighted the influence of human variability on these signals. Compared to Fig. 2, the signals portrayed in Fig. 6 exhibited more variations, posing greater challenges for object recognition.

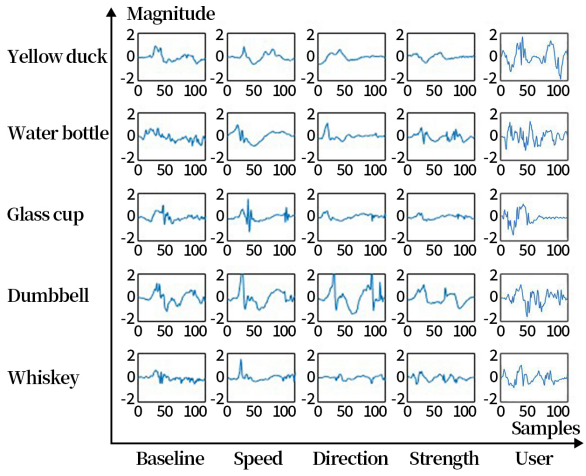

Fig. 6. Example signals of grab-induced vibrations from five daily objects. The initial example corresponds to Fig. 2, while the subsequent ones vary in terms of grasp speeds, positions, strengths, and users.

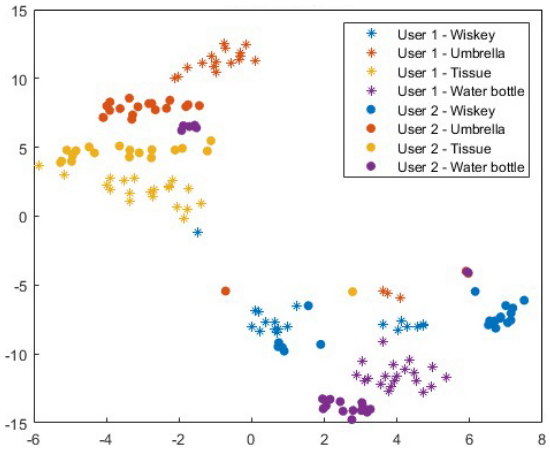

Fig. 7. Interference of subject variability in T-SNE visualization for object differences.

Furthermore, we explored the intricate relationship between human variability and object-specific data representations using T-SNE visualization. As depicted in Fig. 7, the visualization showcased distinct data distributions for different objects, revealing unique signatures associated with each item. Yet, the most striking observation was the diverse data distributions among users interacting with these objects. This variability introduced a complex layer of interference, complicating the clear isolation of object-specific patterns within the data. Despite the challenge posed by individual differences in interactions, a fortuitous discovery emerged: the variations in data among users were smaller than those observed between different objects. This indicated that while users interacted diversely, the distinctions between objects were more prominent. This encouraging finding suggested the feasibility of identifying objects despite individual variability, emphasizing the substantial differences inherent in object data over user interactions. Thus, it offered promise for practical object recognition across diverse users, irrespective of their unique interaction styles.

These experiments shed light on the intricate interplay between human interaction and object-specific vibrations, underscoring the necessity to consider subject variability when studying and interpreting such signals. These findings have paved the way for further exploration, prompting us to refine methodologies for extracting and distinguishing object-based vibrations amidst the inherent variability introduced by human users.

3.7 Key Insights

Based on the aforementioned studies, it is established that the vibration patterns of various objects are distinctive [24, 46]. Additionally, the vibrations induced by object grasping are uniquely identifiable. Our control group study revealed that the system primarily relies on object size and shape for recognition, though material composition also influences identification. Consequently, objects of the same size and shape exhibit varying grasp vibrations due to material disparities, despite identical hand gesture shapes.

Moreover, we conducted an analysis of system variations, including differences in grasp speeds, strengths, directions, and users. Subsequently, in the following section, we will introduce solutions aimed at addressing these challenges.

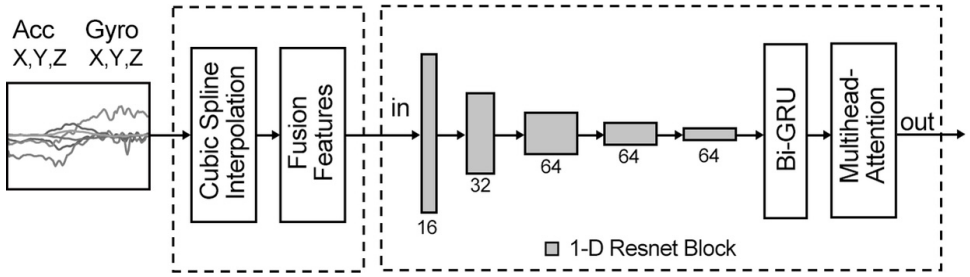

Fig. 8. Feature Extraction.

4 VIOBJECT

4.1 Interference Elimination

4.1.1 Eliminate motion Noise. Although the motion sensor can detect grabbing-object-induced vibrations, it is very sensitive to the user's motions. Fortunately, based on our observations, the frequency spectrum of hand movements is lower than that of grab-induced vibrations. We utilize a 32 Hz Butterworth high pass filter as this is robust to noisy hand movements. [60]. However, grab-induced IMU signals are still too weak to be segmented. ViObject combines 6-axes signals to improve the SNR (Signal Noise Ratio) so that the location of grab-induced vibrations can be identified and segmented. Next, our system uses an energy-based threshold approach [41] to detect the start point of the grab-induced vibration signals. The lower threshold is $\mu + \sigma$, and the higher one is $\mu + 3\sigma$, where μ and σ are the mean and the standard deviation of energy obtained from collected signals, respectively. To calculate the energy, we set the frame length as 0.1s, the frame-shift as 0.01s. The energy threshold is 0.03. Also, if a segmented IMU signal is longer than 2s or shorter than 0.2s, we will ignore this signal as it is noise. After identifying the location of grab-induced vibrations, we utilize GCC (general cross-correlation) based algorithm [36], which is shown in Equation 2 to align signals.

$$r_{xy}(m) = \sum_{n=-\infty}^{\infty} y(n)x(n-m) \tag{2}$$

Where $r_{xy}(m)$ is the cross-correlation function, x and y are two signals from the same class, and m is the parameter of time shift.

4.1.2 Reducing Interference from speeds and strengths. In order to reduce the influence of different grabbing speeds and strengths, we employ three time series data augmentation techniques, as proposed by [60], using all seven combinations resulting from $(2^3 - 1)$. This results in the generation of grab-induced vibrations, increasing the data size by a factor of eight. The three augmentation techniques used are as follows: 1) Zooming, which allows us to simulate different grasp speeds by randomly selecting values from $\times 0.9$ to $\times 1$; 2) Scaling, which allows us to simulate different grasp strengths, with the scaling factor s drawn from a normal distribution with mean 1 and variance 0.2^2, such that $s \in [0, 2]$; 3) Time-warping, which allows us to simulate grasp temporal variance by introducing 2 interpolation knots and warping randomness w drawn from a normal distribution with mean 1 and variance 0.05^2, such that $w \in [0, 2]$.

4.2 Fine-grained Feature Extraction

4.2.1 Cubic Spline Interpolation. Most smartwatches have an IMU sampling rate of 100 Hz, which is too low to accurately capture the high-frequency signals generated by grab-induced vibrations. Taprint [4] and ViBand [39] modified the Linux kernel to improve the IMU sampling rate, but this approach is not widely applicable as Linux

kernels for smartwatches vary by brand and may not be publicly accessible. To overcome this limitation, we use cubic spline interpolation [45] to upsample the sensor readings from 100 Hz to 500 Hz. Cubic spline interpolation is a mathematical technique that constructs a smooth curve that passes through the data points, allowing us to estimate the intermediate values at higher sampling rates. Specifically, given a set of data points (t_i, x_i) where x_i represents the sensor reading at time t_i, we can use cubic spline interpolation to obtain the estimated value $\hat{x}(t)$ at a new time t by solving a set of linear equations that depend on the values of neighboring data points.

Then, we extract amplitude and frequencies from raw vibration signals and fuse them as the inputs to train the classification model. Specifically, we choose power spectral density (PSD) of the collected vibration signals, which reveals the power distribution in different frequencies. If k_i is the received vibrations signals, then the PSD can be defined as

$$PSD_i = 10 \log_{10} \frac{(abs(FFT(k_i))^2}{f_s \times n}, \tag{3}$$

where $FFT(\cdot)$ is the fast Fourier transform operation, f_s is the sampling rate, and n is the number of samples of received signal k_i.

4.2.2 Multihead-Attention Residual Network. In our proposed approach (Fig. 8), the fusion features are first inputted into a series of five Residual Network (ResNet) blocks. [31] These ResNet blocks use skip connections to enable the training of very deep networks, helping to address the problem of vanishing gradients that can arise when training deep neural networks. After each ResNet block, a maximum pooling layer is connected, which reduces the length of the sequence by half.

The ResNet blocks we use in our approach have a kernel size of 1x3. Furthermore, the number of channels in each of the five ResNet blocks is set to 16, 32, 64, 64, and 64, respectively. This configuration is based on empirical studies that have shown that these channel numbers can provide a good balance between the expressiveness of the model and the computational efficiency of training.

After the ResNet blocks, the output is connected to a Bidirectional Gated Recurrent Unit (Bi-GRU) and a Multihead-Attention mechanism. Bi-GRU is a type of recurrent neural network that has been shown to be highly effective for processing sequential data. Our Bi-GRU has a hidden size of 64 and four layers, which provides a good trade-off between expressiveness and computational efficiency.

Multihead-Attention is a mechanism that allows the model to attend to different parts of the input sequence with different weightings. Our Multihead-Attention has an embedding size of 128 and 16 heads. The number of heads determines the number of parallel attention mechanisms that are used to compute the attention scores, while the embedding size determines the dimensionality of the attention output. Specifically, the output of Multihead-Attention can be expressed as:

$$\text{Multihead}(Q, K, V) = \text{Concat}(head_1, head_2, ..., head_h)W^O \tag{4}$$

where Q, K, and V are the query, key, and value vectors of the input sequence, respectively; h is the number of heads; $head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ is the attention output of the i-th head; W_i^Q, W_i^K, and W_i^V are learnable parameters; and W^O is a learnable parameter used to linearly transform the concatenated attention outputs.

Together, these components form a powerful deep learning architecture for processing sequential data to extract features of object identification.

4.3 Combating Orientation Change Impact

In our approach to address the problem of varying vibration responses induced by grabbing different positions of an object from different directions, we use two techniques: center loss and Adversarial Training Regularization.

Center loss is a technique used to improve the clustering of feature representations in a deep neural network. It does this by penalizing the distance between the features and their corresponding class centers, which encourages the network to learn more discriminative features that are closely associated with their corresponding classes. The center loss can be expressed as:

$$L_c = \frac{1}{2}\sum_{i=1}^{N} ||f_i - c_{y_i}||_2^2 \tag{5}$$

where N is the number of training samples, f_i is the feature vector of the i-th sample, c_{y_i} is the class center of the i-th sample's class, and y_i is the class label of the i-th sample. The center loss is added to the overall loss function of the network, and is weighted by a hyperparameter λ to control its contribution to the total loss.

Adversarial Training Regularization is a technique used to improve the robustness of deep neural networks to adversarial attacks. Adversarial attacks are a type of attack where an attacker deliberately introduces small perturbations to the input data in order to fool the network into misclassifying it. Adversarial Training Regularization works by training the network to be robust to such attacks by incorporating adversarial examples into the training process. Specifically, the network is trained on both the original examples and adversarial examples generated by adding small perturbations to the input data. The objective function for Adversarial Training Regularization can be expressed as:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i) + \alpha \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i + \delta_i), y_i) \tag{6}$$

where θ are the parameters of the network, f_θ is the network function, x_i is the i-th training example, y_i is its corresponding label, $\mathcal{L}$ is the loss function, α is a hyperparameter that controls the strength of the adversarial training, and δ_i is the perturbation added to the i-th example to generate its corresponding adversarial example.

Center loss and Adversarial Training Regularization are used together to enhance the robustness and discriminability of deep neural networks for the problem of grabbing different parts of objects from different directions.

4.4 Model Customization

In our object classification system, we recognize the importance of allowing users to customize the classification of their own objects. Users may have unique objects on their desks that are not included in our pre-defined object categories. Furthermore, they may want to add new objects to the system in the future. To address this problem, we use Generalized Few-Shot Learning (FSL) with data synthesis.

4.4.1 Data Synthesis. Although data augmentation techniques can generate signals with larger variances from the data recorded by end-users, these augmented data may not fully capture the actual grabbing fashion variances. To address this issue, we synthesize additional data that simulates the grab-object vibration variance using both the raw signals and the augmented signals as a starting point. We utilize a Δ-encoder, a self-supervised encoder-decoder model [60] that captures the difference between two samples of the same object. The Δ-encoder takes two samples from the same class, creates a small Δ-vector encoding the difference between them, and then uses the vector and one of the samples to reconstruct the other. We trained the Δ-encoder on data from pre-existing objects, drawing pairs of samples randomly to capture within-user variance. In real-time, when the customized a new object's data, we use the Δ-encoder to generate extra samples of the customized objects that contain more natural grabbing variances, enlarging data by 10 times.

4.4.2 Generalized FSL. Generalized FSL is a type of machine learning technique that enables a model to learn to classify new objects based on only a few examples of each class. It achieves this by learning a meta-learning

Fig. 9. Graphical User Interface of ViObject. (a) Customization mode; (b) Modes transition; (3) Prediction mode.

process that can quickly adapt to new classes with very few examples. In our implementation of Generalized FSL, we use a variant of the Prototypical Network architecture.

The Prototypical Network learns a mapping from the input data space to an embedding space where the distance between embeddings corresponds to the similarity between input examples. The embeddings are then used to compute the prototypes of each class, which are used to classify new examples. The objective function for the Prototypical Network can be expressed as:

$$\mathcal{L} = -\sum_{(x,y)\in\mathcal{D}} \log p(y|x) \tag{7}$$

$$= -\sum_{(x,y)\in\mathcal{D}} \log \frac{\exp(-d(f(x), c_y))}{\sum_{y'\in\mathcal{Y}} \exp(-d(f(x), c_{y'}))} \tag{8}$$

where $\mathcal{D}$ is the training set, x is an input example, y is its corresponding label, f is the embedding function, c_y is the prototype of class y, d is a distance metric, and $\mathcal{Y}$ is the set of all classes.

To adapt the Prototypical Network to handle new classes with few examples, we use an additional meta-learning step that learns a function to map from the few-shot examples of new classes to the prototypes of those classes. Meta-learning function to map few-shot examples to prototypes can be expressed as:

$$f_{\theta'}(\{x_i^{(s)}, y_i^{(s)}\}_{i=1}^{K}) \approx c_y \tag{9}$$

where θ' are the meta-learning parameters, $x_i^{(s)}$ and $y_i^{(s)}$ are the input and label of the i-th example from the support set, K is the number of examples in the support set, and c_y is the prototype of the new class. This function can then be used to classify new examples from those classes.

Overall, our use of Generalized FSL enables our system to be highly customizable and adaptable to new objects, while still maintaining high accuracy in object classification.

5 IMPLEMENTATION

We have implemented ViObject both offline and online on commodity Android smartwatch, the ASUS ZenWatch 2 and Huawei Watch 2. ViObject utilizes the built-in accelerometer and gyroscope and acquires the sensor readings through existing Android Wear APIs to detect the grab-object-induced vibrations. The sampling rate through the

APIs for two watch is 200 Hz and 100 Hz, respectively. For the offline system, we have made a data collection app on the smartwatch, then transferred this data to the PC where we used MATLAB to analyze the data. The laptop we used has Intel Core i7 CPU (1.8 GHz) and 16 GB RAM. For the online system, we build an end-to-end standalone application program on a commodity Android smartwatch. We have implemented all the components of our system including signal detection, signal processing, and the neural network algorithm on the smartwatch. As shown in Fig. 9, a user interface was designed to guide the user to customize the model and recognize different objects. In the customization mode, the screen of the smartwatch shows the object's name and the number of times for grabbing an object to customize the model, as shown in Fig. 9 (a). Fig. 9 (b) shows the transition between customization mode and prediction mode. In the prediction mode, the screen shows the prediction of an object's name when the user grabs an object, as shown in Fig. 9 (c). We also implemented Android TextToSpeech [1] to play the sound of the object's name.

For the current on-device (smartwatch) implementation, the app size is 34.7 MB and the average end-to-end latency on the smartwatch is about one second from grabbing the object to displaying the object name. We measure the power consumption of smartwatch using "Battery Historian" from Google. Specifically, three states are measured: 1) idle with the display on, 2) ViObject with power on, but without users grabbing objects, and 3) ViObject with power on with users continuously grabbing objects. Since the platform is only able to measure the percentage of the battery consumption, we record the time duration for consuming 1% battery for each state. The average resulting time duration in each state is 215, 191, and 179 seconds respectively. Given the battery capacity and the working voltage, we calculate the average resulting power consumption of each state, which is 247 mW, 285 mW, and 295 mW respectively. Thus, ViObject only consumes an additional 48 mW of power on top of the base power consumption. For comparison, we also conduct the measurement when running a step counting application, resulting in the power consumption of 288 mW. Thus, the power consumption of ViObject is similar to the typical application running on a smartwatch.

6 EVALUATION

In this section, we first evaluate the performance of our model using a dataset collected from 20 participants and 20 objects. Next, we assess the performance of the customized model for 10 new users and five new objects in Study 2. In Study 3, we evaluate the robustness of the system under various conditions, including different object-touching positions, grasp speed, pressure, times, and smartwatches. We also compare its accuracy to that of traditional machine learning models. Finally, in Study 4, we analyze the usability of the real-time system. The study was approved by the Institutional Review Board (IRB # removed for anonymity).

6.1 Study 1: Pre-trained Model

In this section, we assess the performance of our model using a dataset of 20 objects collected from 20 participants. We begin by presenting the accuracy of each object, including the negative (noise) data. Next, we provide the accuracy of each participant.

6.1.1 Data Collection. We selected 20 common everyday objects as shown in Fig. 10. Examples include whiskey, recognition of which could assist in alcohol addiction applications, and dumbbells, useful for fitness tracking. Even well-established object recognition methods like cameras struggle to identify the vast array of objects in the world. This paper seeks to propose a new concept, and we believe that 20 objects are sufficient to prove this concept and many applications would involve 20 or less objects. We anticipate that 20 objects will suffice for numerous applications (see section 7), such as pill reminders, office desk objects recognition for virtual meetings, and monitoring Alzheimer's disease progression (requiring a few objects). Future research can investigate applications that extend beyond 20 objects. These 20 everyday objects have complex shapes, sizes and material compositions. They represent daily objects that can be used for tangible interaction with natural affordance. Some of them are

Fig. 10. Everyday objects tested in the evaluation: (0) camera (1) small Whisky (2) perfume (3) umbrella (4) Coca-Cola (5) sunglasses (6) a package of tissue (7) insect spray (8) water bottle (9) toy ball (10) sunscreen spray (11) mobile phone (12) glass cup (13) yellow duck (14) dumbbell (15) microphone (16) Bulbasaur (17) skipping rope (18) tapeline (19) baseball player statue.

rigid and heavy (perfume, mobile phone, dumbbell, Coca-Cola, and camera), rigid and light (sunscreen spray, sunglasses, small Whisky and glass cup), heterogeneous (skipping rope, umbrella, toy ball, microphone and tapeline), homogeneous (three small rubber toys: Bulbasaur, yellow duck and baseball player), plastic (water bottle and insect spray) and a package of tissue.

We recruited 20 right-handed participants (12 of them are male) in the age range between [18, 48], which had a mean wrist diameter of 6.8 cm (SD=1.6). Their body mass indexes (BMIs) range from 18.22 (lean) to 29.17 (obese). The BMI illustrates the diversity of users (ranging from thin to overweight individuals). These participants were asked to wear ASUS ZenWatch 2 in a comfortable manner. To demonstrate the basic performance of ViObject, we randomized object order and then asked participants to grab 20 everyday objects in random order. In order to cover the variations in the natural setting, we asked them to grab the objects with different speeds and strengths and grab different parts of the objects from different directions. Users were guided to exercise a minimum of three unique speeds and strengths. Regarding the grasp positions on the object, users grab the left, right, top, and front facets of the object during the interaction. Each object was grabbed by a user 100 times. Thus, we have a sample set with 20 objects × 20 participants × 100 times = 40,000 grabbing-object-induced vibration samples in total. Then, we asked participants to move their hands, wave hands, grab air, applaud, and hit tables randomly for 15 minutes. We took this data as noise. We then transferred this dataset to the PC where we used Matlab to analyze the data.

6.1.2 Results. In order to train and test the model accuracy, we separate the samples of each object in the 3:1:1 proportion for the training set, validation set and test set, respectively. We only present the accuracy of the test set below. We trained a general model from all objects and all participants. Fig. 11 shows the confusion matrix for object classification of 20 daily objects. The cross-validation accuracy for the test set is 86.4% on average (SD=6.08).

0 Camera	90.5	0	0	1.7	0	2.9	0	0.8	0	0.1	0	0.1	0	0	0.1	0.6	0.2	0.2	1.6	0	1.2
1 Samll whisky	0	93.5	4.5	0	0.5	0	0	0.6	0	0	0.3	0	0.1	0	0.1	0	0	0	0	0	0.4
2 Perfume	0	7	83	0	1.4	1.5	0	2.6	0	0	0.2	0	3.3	0	0	0	0	0	0	1	0
3 Umbrella	1.5	0	0	85.5	0	4.4	0.1	0	0	0	0	0.9	0	0	1	1.5	1.4	2	0.5	0.8	0.4
4 Coca-Cola	0	2.2	0.8	0	86.5	0	0.4	1.1	3.4	0	4	0	0.9	0	0	0	0.1	0.1	0.1	0.4	0
5 Sunglasses	0.2	0	0	2.9	0	82.5	0.2	0	0	0.2	0	2.8	0	1.7	0.8	0	2.7	1.1	2.2	2.7	0
6 A package of tissue	0	0	0	1.5	1	0	83	0	2	3.5	0	1	0	0	1.5	1	1	2.5	0	2	0
7 Insect spray	0	3.6	0.1	0	2.1	0	0.1	84.5	0.2	0	6.3	0	2.9	0	0	0	0	0	0	0.2	0
8 Water bottle	0	0.1	0	0	1.5	0	0	0.2	95	1	2.1	0	0	0.1	0	0	0	0	0	0	0
9 Toy ball	0.4	0	0	0.1	0.1	0.4	2.1	0	0.4	86	0	3	0	0	0.8	1.7	1.2	0	0.3	3.4	0.1
10 Sunscreen spray	0	1.5	0.1	0	1.6	0	0	4.4	0.1	0	91.5	0	0.7	0.1	0	0	0	0	0	0	0
11 Mobile phone	0	0	0	2	0	3.4	0	0	0	0	0	84	0	1.8	0.8	0.2	2.5	2	2.6	0.7	0
12 Glass cup	0	0	1.4	0	0	0	0	1.1	0	0.1	0.7	0	96.5	0	0	0.1	0	0	0	0.1	0
13 Yellow duck	0	0	0	0	0.1	1.6	0.1	0.1	0	0.4	0	0.2	0	90	0.3	0.1	4.2	0.1	1.7	0.1	0
14 Dumbbell	0.1	0	0	3.6	0	1.9	0.5	0	0	1.3	0	1.5	0	0.7	80.5	0.3	0.7	3.2	1.7	4	0
15 Microphone	1.1	0.1	0.1	0	0	0.1	0.7	0.1	0	2.5	0	0.1	0	1.6	0.1	91	0.3	0.7	1.4	0.1	0
16 Bulbasaur	0	0	0.1	0.5	0	1.3	0.4	0	0	0.1	0	0.4	0	4.5	1.1	0.6	82.5	1.9	4.2	2.4	0
17 Skipping rope	0.1	0	0	0.6	0	1.4	0	0	0	0.3	0	2.8	0	0.4	4.5	1	1.6	79	5.9	2.4	0
18 Tapeline	0.1	0	0	0.6	0	1.5	0	0	0	0.3	0	6.6	0	4.9	1.9	0	2.1	1.6	77.5	2.9	0
19 Baseball player statue	0.1	0.1	0	2.8	0	0.1	1.4	0.1	0	2.8	0	2.4	1.1	3.7	0.1	1.9	1.7	0.3	3.6	77.5	0.3
20 Negative	0	0	0	0	0	0.1	0	0	0	0	0	1.3	0	0.1	0	0	0	0.1	1.4	0	97

Fig. 11. Confusion matrix of 20 objects. Rows are actual classes and columns are predicted classes.

We observed that objects 13 and 16 were easily confused with each other. We believe that it is because they were similar to each other: Yellow duck (object 13) and Bulbasaur (object 16) are both small rubber toys, with similar materials, sizes and weights. However, the accuracy of the yellow duck (90%) and Bulbasaur (82.5) was good because they have different shapes. Small whisky (object 1) and perfume (object 2) were also sometimes confused with each other since they were both glass bottles with the same materials. However, Small whisky (93.5%) and perfume (83%) also had good accuracy because they have different sizes. Note that detecting the noise had a high accuracy of 97%. We believe that the high accuracy is because grab-object-induced vibrations are much different from noise.

To assess the performance of different participants, we analyzed the accuracy of each individual, as shown in Fig. 12. We found that all participants achieved an accuracy rate above 81%, with User 11 performing the best at 94%. The user demonstrating the highest accuracy is anticipated, as he is most familiar with the system. It should be noted that the test participant still had their training data included in the training set. However, in real-world scenarios, end-users may have limited training data or may be dealing with new objects. To test the system's performance in this context, we conducted a second study.

6.2 Study 2: New Users And New Objects

In this study, we evaluate the performance of our customized model with new users and new objects. Additionally, we investigate the number of times a user must grasp a new object to achieve high accuracy.

6.2.1 Data Collection. We recruited 10 new users (6 males) in the age range between [18, 35], which had a mean wrist diameter of 6.7 cm (SD=1.5). Their body mass indexes (BMIs) range from 17.12 (lean) to 28.87 (obese).

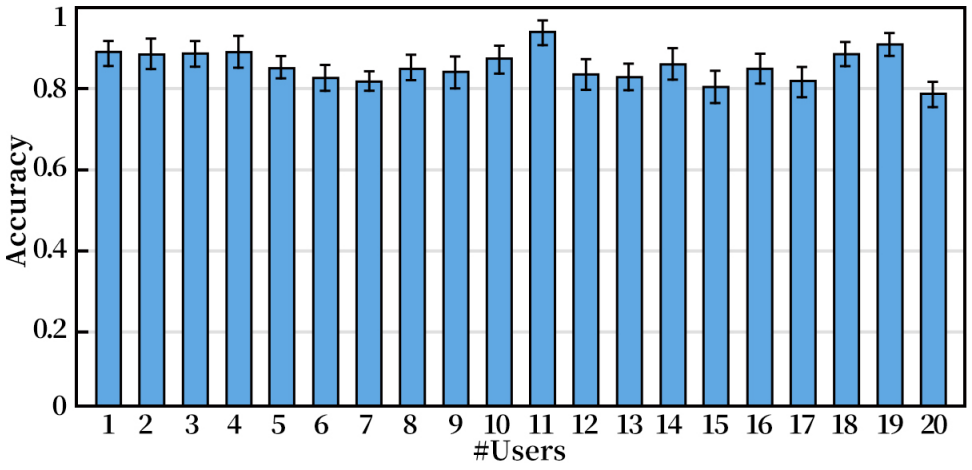

Fig. 12. Accuracies of each user

Fig. 13. Five new objects: a mug, water bottle, computer mouse, book, and pen.

None of the participants participated in study 1. These participants were asked to wear ASUS ZenWatch 2 in a comfortable manner. They have five new objects on their desks, including a mug, water bottle, computer mouse, book, and pen. As shown in Fig. 13, these items are very common daily objects and are very different from the previously illustrated set of 20 objects used for the pretrained model. We asked them to grab each of the objects 30 times from different directions with different speeds and strengths in a random order based on the smartwatch instructions.

6.2.2 Results. The cross-subject (new user) accuracy without any customization (0 shot) stands at 66% as depicted in Fig. 14. To achieve this accuracy, we employ the pretrained model directly for extracting embeddings without any additional training. Subsequently, we calculate the cosine similarity. However, this level of accuracy falls short for effective object recognition. The experimental results demonstrate that object recognition through grab-object induced vibration can be customized for new users and new items using data synthesis and generalized FSL. Fig. 14 show that the 1-shot accuracy is 68.95% (SD: 14.4), 2-shot accuracy is 83.26% (SD: 9.1), 3-shot accuracy is 85.07% (SD: 8.4), 4-shot accuracy is 85.47% (SD: 8.4), and 5-shot accuracy is 90.08% (SD: 8.1).

This experiment shows that ViObject can effectively adapt to new users and items by utilizing a small number of samples for training. The results indicate that the accuracy of the recognition system improves as the number of training samples increases.

In conclusion, this experiment demonstrates that object recognition through grab-object-induced vibrations can be customized for new users and for new items with high accuracy and consistency by grabbing only five times.

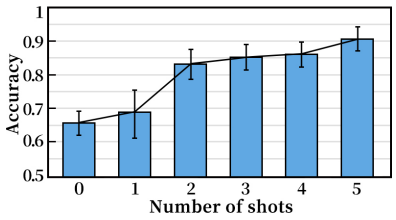

Fig. 14. Customization results with the number of grabbing samples (0-5 shots).

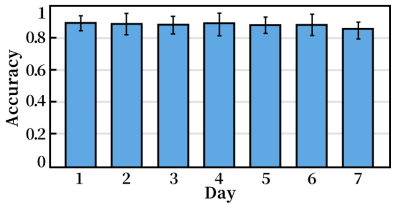

Fig. 15. Performance comparison over seven days.

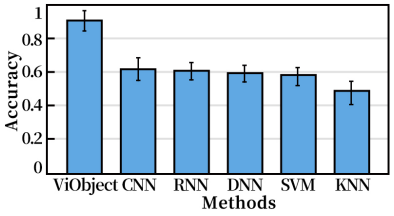

Fig. 16. Performance comparison from other methods.

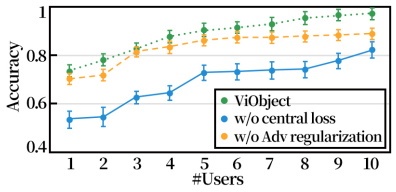

Fig. 17. Ablation study for different object-touching positions.

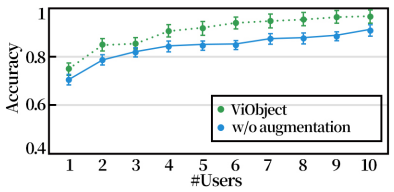

Fig. 18. Ablation study for different grab-object speeds.

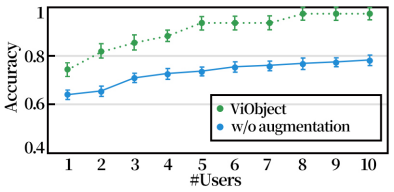

Fig. 19. Ablation study for different grab-object strengths.

6.3 Study 3: Robustness

In order to test the system's robustness under different disturbances, we asked the same 10 participants from Study 2 to conduct a follow-up study. This study included variations in object-touching positions, grasp speeds, strengths, different days, and different smartwatches. Users were guided to exercise a minimum of three unique speeds and strengths. While testing the grasp positions on the object, users grab the left, right, top, and front facets of the object during the interaction. We compared the performance of our model to other traditional machine learning methods, using the data and five-shot model from Study 2.

6.3.1 Different Object-touching Positions. We conducted this experiment to evaluate the accuracy when participants grabbed different parts of the objects. We asked the participants to grab different parts of each object 30 times for testing. Specifically, Users were asked to grab the left, right, top, and front facets of the object during the interaction.

The results of the experiment are summarized in the Fig. 17, where the accuracy of our system is compared with and without adversarial training regularization and center loss. The results show that object recognition through grab-object induced vibration with center loss and adversarial training regularization achieves higher accuracy (87.45%) compared to recognition without them (82.92% without adversarial training regularization and 68.73% without center loss).

6.3.2 Different Speeds. We conducted an experiment to evaluate the accuracy of object recognition through grab-object induced vibration using data augmentation when participants grabbed objects at different speeds. Users were guided to exercise a minimum of three unique speeds. We asked the participants to grab 30 times of each object with different speeds while the vibration signals were recorded.

The results of the experiment are summarized in Fig. 18, where the accuracy of the system is compared with and without data augmentation. The results show that object recognition through grab-object induced vibration with data augmentation achieves higher accuracy (90.07%) compared to recognition without data augmentation (84.30%). The standard deviations also indicate that the performance of the system is consistent across different

users. This is because we used data augmentation techniques during training to create additional samples by varying the grabbing speed.

6.3.3 Different Strengths. We conducted this experiment to evaluate the performance of object recognition through grab-object-induced vibration when participants grabbed objects with different strengths. Users were guided to exercise a minimum of three unique strengths. We asked the participants to grab each object 30 times with varying strengths while the vibration signals were recorded.

The results of the experiment are shown in Fig. 19, where the accuracy of the recognition system is compared with and without data augmentation. The results show that object recognition through grab-object induced vibration achieves higher accuracy (87.45%) when using data augmentation compared to recognition without data augmentation (72.94%). This suggests that data augmentation can help the recognition system adapt to variations in grabbing strength.

6.3.4 Performance over Time. It is crucial to verify that ViObject maintains the temporal stability (i.e., the model of a user is customized once and the resultant system keeps operating in a stable manner over time). In this experiment, we asked participants to grab each object in random order 30 times everyday within one week. Participants only customized the models once on the first day. We tracked the performance of ViObject over time (Fig. 15). The object recognition accuracies for seven days were 89.8%, 88.6%, 87.9%, 88.9%, 87.1%, 87.2%, and 85.1%. The results showed that the accuracy remained constant over one week.

6.3.5 Different smartwatches. Additionally, we asked participants to grab each object 30 times with another smartwatch: Huawei Watch 2. Compared to ASUS Zenwatch 2, the new watch also achieved similar accuracy of 89.9%. We believe that different types of IMUs should not impact the model performances.

6.3.6 Comparing to Other Methods. The results of our experiments demonstrate that object recognition through grab-object induced vibration using our proposed system achieved the best accuracy (90.08%) compared to other traditional machine learning methods such as CNN (61.3%), RNN (60.5%), DNN (59.4%), SVM (58.1%), and KNN (48.8%). (Fig. 16) This indicates that our system is highly effective in recognizing objects based on their vibration patterns. The superior performance of our system may be attributed to the use of data augmentation, attention mechanism, center loss, adversarial training regularization, and data synthesis, which can enhance the system's ability to recognize objects accurately even in the face of variations in grabbing speed, strength, and location. Overall, our experiments demonstrate that object recognition through grab-object-induced vibration using our proposed system can achieve high accuracy.

6.4 Study 4: Usability

Using the data collected in Study 1, we developed a pre-trained model based on the data from 20 participants. Building on the findings of Study 1, we then created ViObject as an end-to-end real-time system for a commodity Android smartwatch (ASUS ZenWatch 2). For the usability study, we recruited volunteers to use the system and complete the System Usability Scale (SUS) [2] and Task Load Index (NASA-TLX) questionnaire [30].

6.4.1 Procedure. We recruited an additional ten participants, six of whom were male, aged between 18 and 38 years. They had a mean wrist diameter of 6.3 cm (SD = 1.1), and their body mass indexes (BMIs) ranged from 19.31 (lean) to 28.5 (obese). Participants were asked to wear the smartwatch comfortably.

Initially, we instructed users to activate the customization mode in the app. The smartwatch's screen displayed the name of an item and a number indicating how many times the object should be grabbed, as shown in Fig. 9 (a). As users grabbed objects, the number on the smartwatch changed to reflect the remaining grabs. Once all objects were customized, the customization mode ended. If the app mistook motion noise for an object grab-induced

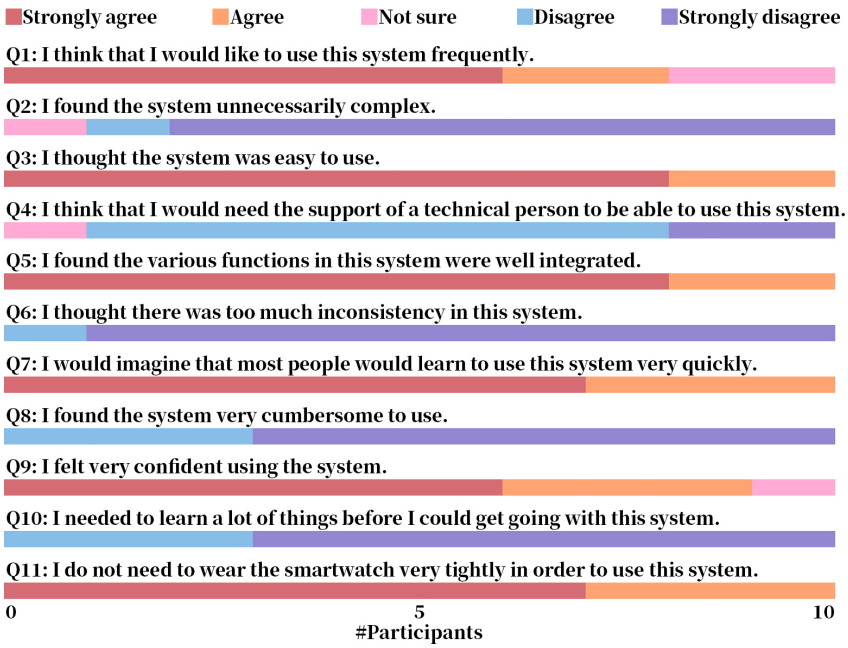

Fig. 20. SUS based standard user study result.

vibration signal, users could click the "DELETE" button; however, we counted this as a false positive. If users grabbed an object and the app did not react, they could grab the object again, and we counted it as a false negative.

Next, we asked users to click the testing button. The app then customized the model using the personal data. Following this process, the app entered testing mode. Users could grab objects, and the screen displayed the name of the predicted object. We requested that each user randomly select five objects in their house for object identification and spend a whole day with the app. Finally, participants completed the System Usability Scale (SUS) [2] and Task Load Index (NASA-TLX) questionnaire [30].

6.4.2 Results. There are ten questions in the SUS [2]. Additionally, we added another question related to smartwatch wearing: I do not need to wear the smartwatch very tightly in order to use this system. The questionnaire results as shown in Figure 20 show that participants gave positive feedback on the usability of the ViObject system. To be specific, the questionnaire asked questions with five response options for respondents, from Strongly Agree to Strongly Disagree. The questions and the results are as follows: (1) I think that I would like to use this system frequently. (2 Not Sure, 2 Agree, 6 Strongly agree) (2) I found the system unnecessarily complex. (8 Strongly disagree, 1 Disagree, 1 Not Sure) (3) I thought the system was easy to use. (2 Agree, 8 Strongly agree) (4) I think that I would need the support of a technical person to be able to use this system. (2 Strongly disagree, 7 Disagree, 1 Not Sure) (5) I found the various functions in this system were well integrated. (2 Agree, 8 Strongly agree) (6) I thought there was too much inconsistency in this system. (9 Strongly disagree, 1 Disagree) (7) I would imagine that most people would learn to use this system very quickly. (3 Agree, 7 Strongly Agree) (8) I found the system very cumbersome to use. (7 Strongly disagree, 3 Disagree) (9) I felt very confident using the system. (1 Not Sure, 3 Agree, 6 Strongly agree) (10) I needed to learn a lot of things before I could get going with this system. (7 Strongly disagree, 3 Disagree) (11)I do not need to wear the smartwatch very tightly in order to use this system. (3 Agree, 7 Strongly agree) The scores and the results support that ViObject is comfortable, user-friendly, and easy to use.

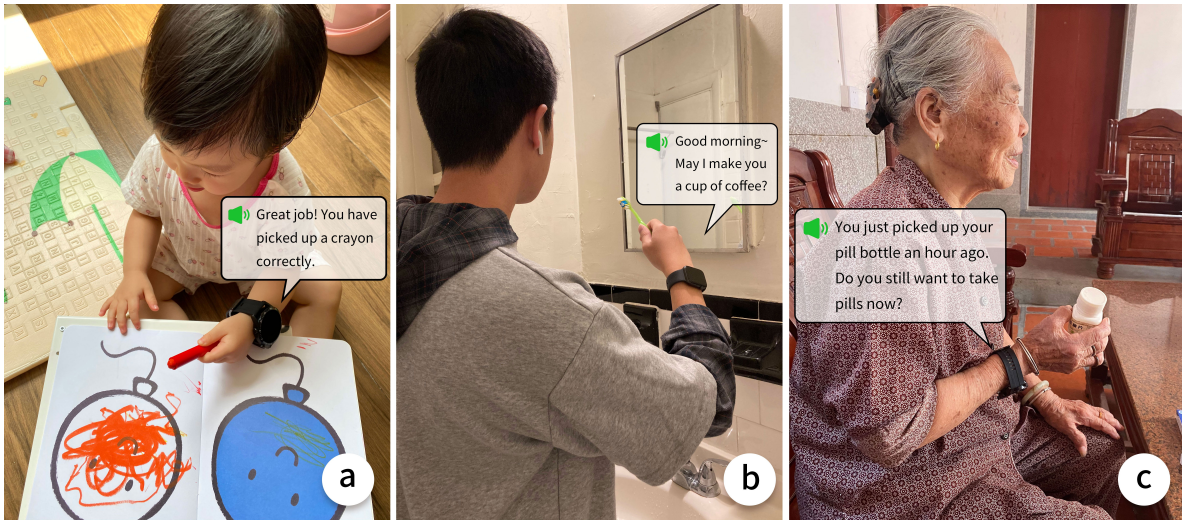

Fig. 21. ViObject applications: (a) Learning assistants (b) Smart services (c) Medication reminders

The NASA-TLX is a tool for measuring and conducting a subjective workload assessment. It rates performance across six dimensions to determine an overall workload rating. Participants were asked to rate their scores on a scale ranging from low (1) to high (7). The six dimensions are as follows: 1. Mental demand: how much thinking, deciding, or calculating was required to perform the task. 2. Physical demand: the amount and intensity of physical activity required to complete the task. 3. Temporal demand: the amount of time pressure involved in completing the task. 4. Performance: the level of success in completing the task. 5. Effort: how hard does the participant have to work to maintain their level of performance? 6. Frustration level: how insecure, discouraged, or secure or content the participant felt during the task. Participants also reported low task load and little frustration from the NASA-TLX questionnaire results (on a 7-point Likert scale): low task mental load (1.3±1.2), low physical load (2.4±1.1), low temporal load (2.4±1.2), low effort (2.5±1.3), little frustration (1.3±0.9) and good performance (6.4±1.1).

Overall, the study suggests that the ViObject system has good usability and promising potential for future use.

7 APPLICATIONS

In this section, we discuss some applications that can benefit from ViObject because of how it interfaces humans and objects seamlessly. While these four applications can be solved by various other techniques, ViObject does not require actions found in these solutions such as attaching a tag on the target[39], taking a picture in front of a target, using an active device to make vibrations[46], or taking out a phone to knock on objects[24]. These necessary extra trigger actions affect our daily activities, separate our physical world and technology world, and are not applicable to some applications, such as medication reminders and smart services. In contrast, ViObject aims to create borderless and fluid interactions between the technological world and our daily life. Simply by recognizing passive vibrations from grabbing objects, ViObject changes the world itself into an interface.

7.1 Education

The use of everyday objects as a control interface can also be applied to educational and entertainment applications, such as interactive exhibits in museums or immersive experiences in theme parks. By using familiar objects as

a control interface, these experiences can be made more engaging and interactive for users. For example, this can assist mentally or visually impaired children as shown in Fig. 21 (a). When children with mental or visual impairments are young, they learn about shapes and textures before eventually learning objects [56]. We believe that ViObject can help these children learn different objects if trained on a large variety of daily objects. In this way, we can teach the child different objects' names and an individual does not have to present for the child to practice or even learn to identify new objects. For example, when our system is activated and a child picks up an umbrella, the smartwatch may spell the word on the watch screen and as an example will say: "This is an umbrella. Do you want me to tell you how to use it?"

7.2 Smart Home Services

Our system can provide smart services by using mundane objects as triggers for specific actions. For example, when a user picks up his/her key, our system asks:" Are you leaving home? Do you want me to turn off lights or turn off appliances?" This helps users make sure that the dangerous appliances are off when they leave home. To give another example as shown in Fig. 21 (b), when a user picks up a toothbrush at the morning, our system asks:" Good morning, Tom. May I make you a cup of coffee?" If Tom chooses yes, our system sends a command to a coffee machine in the kitchen and makes a cup of coffee for Tom when he is still in the bathroom.

7.3 Healthcare

Medication Reminders: Many people find that remembering to take pills is difficult. With our system, users can set a reminder on the smartwatch. For example, if our system does not detect that the user picks up the pill bottle from 2 pm-7 pm, the smartwatch will notify the user. The system can also track whether the user picks up the pill bottle or not after the reminder. For another example as shown in Fig. 21 (c), if our system detects that the user is grabbing a pill bottle, but there is a record on our system that the user had picked up the pill bottle earlier during the day, our system can send warnings on the smartwatch screen and as an example will say:" Hi, Tom. You just picked up your pill bottle an hour ago. Do you still want to take pills now?"

AD Progression Assessments: To assess Alzheimer's disease progression, doctors usually give patients some simple cognitive tasks and observe patients' reactions. However, it is difficult for doctors and patients to meet each other every day or every week to track the progression. With our system, the smartwatch could send these tasks to patients and the results back to the doctors for progression assessment. For instance, our system could ask the patient to grab a water bottle, then ask him/her to pick up a pen, and so on. Our system could track how many times the patient picks up a correct object and how many times the patient picks up a wrong object, and then send the result to the doctor. In this way, our system can facilitate Alzheimer's disease progression assessment.

7.4 Metaverse

ViObject's passive object recognition technology could be used in the metaverse to enhance the user experience by enabling more realistic and interactive virtual environments. For example, a user wearing a smartwatch with ViObject technology could pick up a physical object such as a book, and the metaverse could replicate the book in the virtual environment, allowing the user to interact with it in the same way they would in the real world. This could enhance the immersive experience of the metaverse and allow for more natural interactions with virtual objects. Furthermore, ViObject could be used in the metaverse to enable users to bring their own physical objects into the virtual environment. This could be useful for creating personalized virtual spaces or for using physical tools or devices within a virtual context. Overall, ViObject's passive object recognition technology has the potential to enable more seamless and interactive experiences in the metaverse by bridging the gap between the physical and virtual worlds.

7.5 Entertainment

Objects as a control interface could enhance entertainment applications, including interactive exhibits in museums or immersive experiences in theme parks. This approach can increase user engagement and interactivity. For instance, in escape room games, ViObject can recognize when visitors pick up specific objects and trigger clues or provide additional information through the smartwatch to assist in solving puzzles and progressing through the game.

8 LIMITATION AND FUTURE WORK

Our system exhibits several limitations. Firstly, it is not entirely passive, necessitating customization and training for utilization by new users. Exploring future avenues such as zero-shot learning for seamless integration with new users is a potential solution. Furthermore, variations in how users wear their watches pose a challenge, emphasizing the need for adaptable approaches. Additionally, our retained model is only trained with 20 objects. Addressing these issues is complex due to the constraints of limited training data. Potential remedies involve exploring efficient methods for large dataset collection, virtual data generation, and learning features from diverse modalities. Secondly, the smartwatch's ability to detect a grasping event and identify the grasped object relies on being worn on the dominant or grasping hand. However, this requirement is not commonly met among users of commercial smartwatches. Anticipating a future where users may wear smart wristbands on both hands, adaptations to our system may be needed. Thirdly, individuals with Parkinson's disease may exhibit small movements or shaking in their hands while grabbing or holding objects. A targeted study of this demographic is vital to enhance the accessibility and effectiveness of our system. Fourthly, we will investigate false positives associated with the detection of grasping events in real-world scenarios. Finally, we can integrate contextual information such as hand activities with the object in use. For instance, when we detect a user typing on a keyboard, they are more inclined to reach for a mouse afterward. Incorporating this content information can significantly enhance the system's performance.

9 CONCLUSION

In conclusion, we have introduced ViObject, a novel system for passive object recognition that utilizes accelerometer and gyroscope sensor data from commodity smartwatches to identify untagged everyday objects. Through our design and implementation process, we have successfully addressed challenges such as motion interference, grasp speed/pressure variations, object-touching position changes, and customization for new users and new objects. We have demonstrated the feasibility of passive object recognition using commodity smartwatches and shown promising results in terms of recognition performance and user feedback. ViObject has the potential to revolutionize the way we interact with our physical environment, from smart home automation to healthcare and assistive technologies. With the rapid development of smartwatch technology and its widespread adoption, we believe that ViObject will have a significant impact on the field of human-object interaction in the future.

ACKNOWLEDGMENTS

This work was supported, in part, by the National Science Foundation Research Traineeship Program Grant 1829004. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies of the funding agencies.

REFERENCES

[1] [n.d.]. Android TextToSpeech. https://developer.android.com/reference/android/speech/tts/TextToSpeech.

[2] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[3] Wenqiang Chen, Daniel Bevan, and John Stankovic. 2021. ViObject: A Smartwatch-based Object Recognition System via Vibrations. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. 97–99.

[4] Wenqiang Chen, Lin Chen, Yandao Huang, Xinyu Zhang, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2019. Taprint: Secure text input for commodity smart wristbands. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[5] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Shwetak Patel, and John Stankovic. 2021. ViFin: Harness Passive Vibration to Continuous Micro Finger Writing with a Commodity Smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[6] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Patel Shwetak, and John Stankovic. 2020. Continuous micro finger writing recognition with a commodity smartwatch: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 603–604.

[7] Wenqiang Chen, Lin Chen, Kenneth Wan, and John Stankovic. 2020. A smartwatch product provides on-body tapping gestures recognition: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 589–590.

[8] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2018. Vitype: A cost efficient on-body typing system through vibration. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[9] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2019. A Low Latency On-Body Typing System through Single Vibration Sensor. *IEEE Transactions on Mobile Computing* 19, 11 (2019), 2520–2532.

[10] Wenqiang Chen, Maoning Guan, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2017. FLoc: Device-free passive indoor localization in complex environments. In *2017 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[11] Wenqiang Chen, Yexin Hu, Wei Song, Yingcheng Liu, Antonio Torralba, and Wojciech Matusik. 2024. CAvatar: Real-time Human Activity Mesh Reconstruction via Tactile Carpets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–24.

[12] Wenqiang Chen, Yanming Lian, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2017. Virtual keyboard for wearable wristbands. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–2.

[13] WENQIANG CHEN, SHUPEI LIN, ELIZABETH THOMPSON, and JOHN STANKOVIC. 2021. SenseCollect: We Need Efficient Ways to Collect On-body Sensor-based Human Activity Data! *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* 1, 1 (2021).

[14] Wenqiang Chen and John Stankovic. 2022. ViWatch: harness vibrations for finger interactions with commodity smartwatches. In *Proceedings of the 13th ACM Wireless of the Students, by the Students, and for the Students Workshop*. 4–6.

[15] Wenqiang Chen, Ziqi Wang, Pengrui Quan, Zhencan Peng, Shupei Lin, Mani Srivastava, Wojciech Matusik, and John Stankovic. 2023. Robust Finger Interactions with COTS Smartwatches via Unsupervised Siamese Adaptation. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[16] Wenqiang Chen, Ziqi Wang, Pengrui Quan, Zhencan Peng, Shupei Lin, Mani Srivastava, and John Stankovic. 2022. Making Vibration-based On-body Interaction Robust. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 300–301.

[17] W Chen, J Zhu, J Stankovic, G Lauder, and H Bart-Smith. 2021. Tuna robotics: using machine learning and inertial measurement sensors for sensory feedback during swimming. In *INTEGRATIVE AND COMPARATIVE BIOLOGY*, Vol. 61. OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA, E134–E134.

[18] Jungchan Cho, Inhwan Hwang, and Songhwai Oh. 2016. VibePhone: efficient surface recognition for smartphones using vibration. *Pattern Analysis and Applications* 19, 1 (2016), 251–265.

[19] Adrian A de Freitas, Michael Nebeling, Xiang'Anthony' Chen, Junrui Yang, Akshaye Shreenithi Kirupa Karthikeyan Ranithangam, and Anind K Dey. 2016. Snap-to-it: A user-inspired platform for opportunistic device interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5909–5920.

[20] Viktor Erdélyi, Hamada Rizk, Hirozumi Yamaguchi, and Teruo Higashino. 2021. Learn to see: A microwave-based object recognition system using learning techniques. In *Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking*. 145–150.

[21] Junjun Fan, Xiangmin Fan, Feng Tian, Yang Li, Zitao Liu, Wei Sun, and Hongan Wang. 2018. What is that in your hand? recognizing grasped objects via forearm electromyography sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–24.

[22] Arjan Geven, Peter Strassl, Bernhard Ferro, Manfred Tscheligi, and Harald Schwab. 2007. Experiencing real-world interaction: results from a NFC user experience field trial. In *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*. 234–237.

[23] Jun Gong, Yu Wu, Lei Yan, Teddy Seyed, and Xing-Dong Yang. 2019. Tessutivo: Contextual interactions on interactive fabrics with inductive sensing. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 29–41.

[24] Taesik Gong, Hyunsung Cho, Bowon Lee, and Sung-Ju Lee. 2019. Knocker: Vibroacoustic-based object recognition with smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–21.

[25] Tobias Grosse-Puppendahl, Sebastian Herber, Raphael Wimmer, Frank Englert, Sebastian Beck, Julian Von Wilmsdorff, Reiner Wichert, and Arjan Kuijper. 2014. Capacitive near-field communication for ubiquitous interaction and perception. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 231–242.

[26] Maoning Guan, Wenqiang Chen, Yandao Huang, Rukhsana Ruby, and Kaishun Wu. 2019. FaceInput: a hand-free and secure text entry system through facial vibration. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[27] Chris Harrison and Scott E Hudson. 2008. Lightweight material detection for placement-aware mobile computing. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. 279–282.

[28] Chris Harrison, Desney Tan, and Dan Morris. 2010. Skinput: appropriating the body as an input surface. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–462.

[29] Chris Harrison, Robert Xiao, and Scott Hudson. 2012. Acoustic barcodes: passive, durable and inexpensive notched identification tags. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 563–568.

[30] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[32] Yongzhi Huang, Kaixin Chen, Yandao Huang, Lu Wang, and Kaishun Wu. 2021. Vi-liquid: unknown liquid identification with your smartphone vibration.. In *MobiCom*. 174–187.

[33] Yandao Huang, Wenqiang Chen, Hongjie Chen, Lu Wang, and Kaishun Wu. 2019. G-fall: Device-free and training-free fall detection with geophones. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[34] Hiroshi Ishii and Brygg Ullmer. 1997. Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '97)*. Association for Computing Machinery, New York, NY, USA, 234–241. https://doi.org/10.1145/258549.258715

[35] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. 2017. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5839–5849.

[36] Charles H. Knapp and G. Clifford Carter. 1976. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing* (1976).

[37] Kai Kunze and Paul Lukowicz. 2007. Symbolic object localization through active sampling of acceleration and sound signatures. In *International Conference on Ubiquitous Computing*. Springer, 163–180.

[38] Gierad Laput and Chris Harrison. 2019. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[39] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 321–333.

[40] Gierad Laput, Chouchang Yang, Robert Xiao, Alanson Sample, and Chris Harrison. 2015. Em-sense: Touch recognition of uninstrumented, electrical and electromechanical objects. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 157–166.

[41] Jian Liu, Yingying Chen, Marco Grutese, and Yan Wang. [n.d.]. VibSense: Sensing Touches on Ubiquitous Surfaces through Vibration. In *Proc. IEEE Secon, 2017*.

[42] Jian Liu, Yingying Chen, Marco Gruteser, and Yan Wang. 2017. Vibsense: Sensing touches on ubiquitous surfaces through vibration. In *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[43] Shan Luo, Leqi Zhu, Kaspar Althoefer, and Hongbin Liu. 2017. Knock-knock: acoustic object recognition by using stacked denoising autoencoders. *Neurocomputing* 267 (2017), 18–24.

[44] Takuya Maekawa, Yasue Kishino, Yutaka Yanagisawa, and Yasushi Sakurai. 2012. WristSense: wrist-worn sensor device with camera for daily activity recognition. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 510–512.

[45] Sky McKinley and Megan Levine. 1998. Cubic spline interpolation. *College of the Redwoods* 45, 1 (1998), 1049–1060.

[46] Seungjae Oh, Gyeore Yun, Chaeyong Park, Jinsoo Kim, and Seungmoon Choi. 2019. VibEye: Vibration-Mediated Object Recognition for Tangible Interactive Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[47] Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, and Tatsuya Harada. 2016. Recognizing activities of daily living with a wrist-mounted camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3103–3111.

[48] Julius Cosmo Romeo Rudolph, David Holman, Bruno De Araujo, Ricardo Jota, Daniel Wigdor, and Valkyrie Savage. 2022. Sensing Hand Interactions with Everyday Objects by Profiling Wrist Topography. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 1–14.

[49] Munehiko Sato, Shigeo Yoshida, Alex Olwal, Boxin Shi, Atsushi Hiyama, Tomohiro Tanikawa, Michitaka Hirose, and Ramesh Raskar. 2015. Spectrans: Versatile material classification for interaction with textureless, specular and transparent surfaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2191–2200.

[50] Adwait Sharma, Joan Sol Roo, and Jürgen Steimle. 2019. Grasping microgestures: Eliciting single-hand microgestures for handheld objects. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[51] Lei Shi, Maryam Ashoori, Yunfeng Zhang, and Shiri Azenkot. 2018. Knock knock, what's there: converting passive objects into customizable smart controllers. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–13.

[52] Nicolas Villar, Daniel Cletheroe, Greg Saul, Christian Holz, Tim Regan, Oscar Salandin, Misha Sra, Hui-Shyong Yeo, William Field, and Haiyan Zhang. 2018. Project zanzibar: A portable and flexible tangible interaction platform. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[53] Edward J. Wang, Tien-Jui Lee, Alex Mariakakis, Mayank Goel, Sidhant Gupta, and Shwetak N. Patel. 2015. MagnifiSense: Inferring Device Interaction Using Wrist-Worn Passive Magneto-Inductive Sensors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 15–26. https://doi.org/10.1145/2750858.2804271

[54] Ju Wang, Jie Xiong, Xiaojiang Chen, Hongbo Jiang, Rajesh Krishna Balan, and Dingyi Fang. 2017. TagScan: Simultaneous target imaging and material identification with commodity RFID devices. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 288–300.

[55] Roy Want, Kenneth P Fishkin, Anuj Gujar, and Beverly L Harrison. 1999. Bridging physical and virtual worlds with electronic tags. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 370–377.

[56] Carmen Willings. 2017. Concept Development. https://www.teachingvisuallyimpaired.com/concepts-to-teach.html

[57] Kaishun Wu, Yandao Huang, Wenqiang Chen, Lin Chen, Xinyu Zhang, Lu Wang, and Rukhsana Ruby. 2020. Power saving and secure text input for commodity smart watches. *IEEE Transactions on Mobile Computing* 20, 6 (2020), 2281–2296.

[58] Te-Yen Wu, Lu Tan, Yuji Zhang, Teddy Seyed, and Xing-Dong Yang. 2020. Capacitivo: Contact-Based Object Recognition on Interactive Fabrics using Capacitive Sensing. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 649–661.

[59] Robert Xiao, Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Deus EM Machina: on-touch contextual functionality for smart IoT appliances. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4000–4008.

[60] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, et al. 2022. Enabling hand gesture customization on wrist-worn devices. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[61] Xiangyu Xu, Jiadi Yu, Yingying Chen, Qin Hua, Yanmin Zhu, Yi-Chao Chen, and Minglu Li. 2020. TouchPass: towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[62] Lin Yang, Wei Wang, and Qian Zhang. 2016. VibID: User identification through bio-vibrometry. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 1–12.

[63] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. 2012. Magic finger: always-available input through finger instrumentation. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 147–156.

[64] Hui-Shyong Yeo, Gergely Flamich, Patrick Schrempf, David Harris-Birtill, and Aaron Quigley. 2016. Radarcat: Radar categorization for input & interaction. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 833–841.

[65] Hui-Shyong Yeo, Juyoung Lee, Andrea Bianchi, David Harris-Birtill, and Aaron Quigley. 2017. Specam: Sensing surface color and material with the front-facing camera of a mobile device. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–9.

[66] Neng-Hao Yu, Li-Wei Chan, Seng Yong Lau, Sung-Sheng Tsai, I-Chun Hsiao, Dian-Je Tsai, Fang-I Hsiao, Lung-Pan Cheng, Mike Chen, Polly Huang, et al. 2011. TUIC: enabling tangible interaction on capacitive multi-touch displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2995–3004.

[67] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, et al. 2018. FingerPing: Recognizing fine-grained hand poses using active acoustic on-body sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.

[68] Maotian Zhang, Qian Dai, Panlong Yang, Jie Xiong, Chang Tian, and Chaocan Xiang. 2018. idial: Enabling a virtual dial plate on the hand back for around-device interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–20.

[69] Shibo Zhang, Qiuyang Xu, Sougata Sen, and Nabil Alshurafa. 2020. VibroScale: turning your smartphone into a weighing scale. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 176–179.

[70] Yang Zhang, Gierad Laput, and Chris Harrison. 2018. Vibrosight: Long-range vibrometry for smart environment sensing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 225–236.

[71] Yiran Zhao, Shuochao Yao, Shen Li, Shaohan Hu, Huajie Shao, and Tarek F Abdelzaher. 2017. VibeBin: A vibration-based waste bin level detection system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–22.